On depth based classification of functional data

Sami Helander

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 22.10.2018

Thesis supervisor:

Asst. Prof. Pauliina Ilmonen

Thesis advisor:

Adj. Prof. Lauri Viitasaari

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Sami Helander

Title: On depth based classification of functional data

Date: 22.10.2018 Language: English Number of pages: 5+51

Department of Mathematics and Systems Analysis

Professorship: SCI3053

Supervisor: Asst. Prof. Pauliina Ilmonen

Advisor: Adj. Prof. Lauri Viitasaari

The purpose of this Master's Thesis is to discuss the application of functional statistical depth, a powerful nonparametric modeling tool, to supervised functional classification.

With the recent rapid increase of the sophistication of measurement and storage tools, we have begun to encounter more and more complex datasets on all fields of research. This sudden explosion of very high dimensional complex data has brought with it an increasing need for inferential analytic tools for dealing with such data. However, developing methodology for functional data is far from straightforward due to the introduction of a wide range of important features unique to this type of data, most notably, shape and shape-outlyingness. The issue is furthermore complicated by the massive computational load many otherwise appealing approaches would impose.

In this thesis, shape receptive depth based classification is considered. In particular, the focus is on Jth order kth moment integrated depth based classification.

Receptiveness to shape features and shape-outlyingness of the Jth order kth moment integrated depth is discussed and important key-ideas related to its features are established. Then, the Jth order kth moment integrated depth is applied to supervised functional classification for two different real datasets. Performance of different functional depth approaches is compared. The real data examples illustrate excellent classification accuracy of the Jth order kth moment integrated depth. Finally, future work and improvement suggestions on the area are discussed.

Keywords: functional data analysis, statistical depth, classification, shape, outlyingness

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Sami Helander

Työn nimi: Syvyysmittaan perustuvasta funktionaalisen aineiston luokittelusta

Päivämäärä: 22.10.2018 Kieli: Englanti Sivumäärä: 5+51

Matematiikan ja systeemianalyysin laitos

Professuuri: SCI3053

Työn valvoja: Asst. Prof. Pauliina Ilmonen

Työn ohjaaja: Adj. Prof. Lauri Viitasaari

Työn tavoitteena on tarkastella funktionaalisen tilastollisen syvyyden soveltamista luokittelussa. Syvyysmitat ovat epäparametrisia mittareita, jotka kertovat havaintojen tilastollisesta poikkeavuudesta.

Tänä päivänä pystymme tallentamaan valtavia määriä dataa. Tämä on mahdollistanut monimutkaisten ja korkeauloitteisten aineistojen keräämisen ja analysoinnin. Tästä on syntynyt tarve uusille menetelmille jotka soveltuvat korkeauloitteisen datan käsittelyyn. Funktionaaliset aineistot ovat ääretönuloitteisia. Menetelmien kehittäminen ääretönuloitteisten aineistojen analysointiin on vaikeaa. Erityisen haastavaa on huomioida funktionaalisten havaintojen muoto. Lisäksi laskennallinen taakka saattaa tuottaa ongelmia.

Työssä tutkitaan funktioiden muotoa huomioivien syvyysmittarian käyttöä luokittelussa. Erityisesti, työssä tarkastellaan J:nnen asteen k:nnen momentin integroituun syvyysmittaan perustuvaa luokittelua.

Työssä tarkastellaan J:nnen asteen k:nnen momentin integroidun syvyysmitan herkkyyttä funktioiden muodoille ja muodon suhteen poikkeaville havainnoille. J:nnen asteen k:nnen momentin integroitua syvyysmittaa käytetään kahden oikean aineiston luokitteluun. Menetelmän suorituskykyä verrataan muihin syvyysmittaan perustuviin luokittelijoihin. Aineistoesimerkit havainnollistavat J:nnen asteen k:nnen momentin integroidun syvyysmitan erinomaista luokittelukykyä. Työn lopussa esitetään ajatuksia mahdollisuuksista parantaa ja laajentaa tarkasteltua menetelmää.

Avainsanat: funktionaalinen data-analyysi, tilastollinen syvyys, luokittelu, muoto, tilastollinen poikkeavuus

Preface

I wish to thank my supervisor Pauliina Ilmonen and my advisor Lauri Viitasaari for excellent guidance and interesting discussions.

I also wish to thank my friends and especially my family for their constant support and encouragement in my studies.

Otaniemi, 22.10.2018

Sami Helander

Contents

1 Introduction

Today, we are able to store massive amounts of data, and large dimensional datasets are becoming more and more common. One approach for dealing with very large dimensional data is to assume that the observations are random functions, instead of random vectors. This approach is well justified especially in the cases when the dataset consists of numerous observations of the same process. Analysing the data as functional, instead of large dimensional vectors, enables to apply methods designed for processes of continuous nature.

Thinking of the data as functional means thinking of the observation sequences as single entities that are continuous in time, or some other continuum, over which they are being sampled. In practice however, this often means that instead of functions, the observed functions x_i often arrive to us as sequences of value-index pairs $x_i = (y_{ij}, t_{ij})$, where $j = 1, 2, \ldots, n_i$. This means that, although the functions and processes being measured are continuous, we can never directly observe them in their entirety as this would mean measuring and storing uncountably many values. Instead, each observation is only ever partially observed at a certain set of measurement points which need not even be the same ranging from an observation to another. Thus, continuity of the functions here means that in principle if we chose to, we could measure the process at any arbitrary point in time.

The context of the data or the phenomenon to be analyzed often -but not always- gives raise to an assumption of a certain degree of smoothness on the functions. That is, given a fine enough measurement scale, two adjacently measured function values necessarily depend on eachother, and cannot be arbitrarily far apart. The interplay between smooth and rough plays an important role in the treatment of functional data and is discussed further in Section 2.

As functional data appears more and more commonly in different applications, development of the theory and methods for analysing such data has become more and more important. Many methods originally developed for multivariate data have already seen extensions to the functional context (see for example Ramsay and Silverman (2005)). However, extending multivariate methods to functional setting is not straightforward; The introduction of infinite dimensionality brings about a wide range of features, concepts and difficulties that are not present in finite dimensional cases.

Difficulties in dealing with functional data are not only limited to those brought about by the continuous functional structure of the data. Also, the form in which the data arrives to us can be challenging and might require some ingenuity in how to pre-process and present the data in an informative way. Good examples of such challenging forms the data might arrive in are plentiful in the literature. A common example is data that consists of input-output pairs of two clearly connected functional sources that should be analysed jointly. Also, data where the observations consist of cyclic processes that are continuously observed over multiple cycles is commonly encountered. This type of data also often exhibits trends that span multiple sub cycles, posing additional challenges for the analysis.

In finite dimensions the relative location of an observation with respect to a

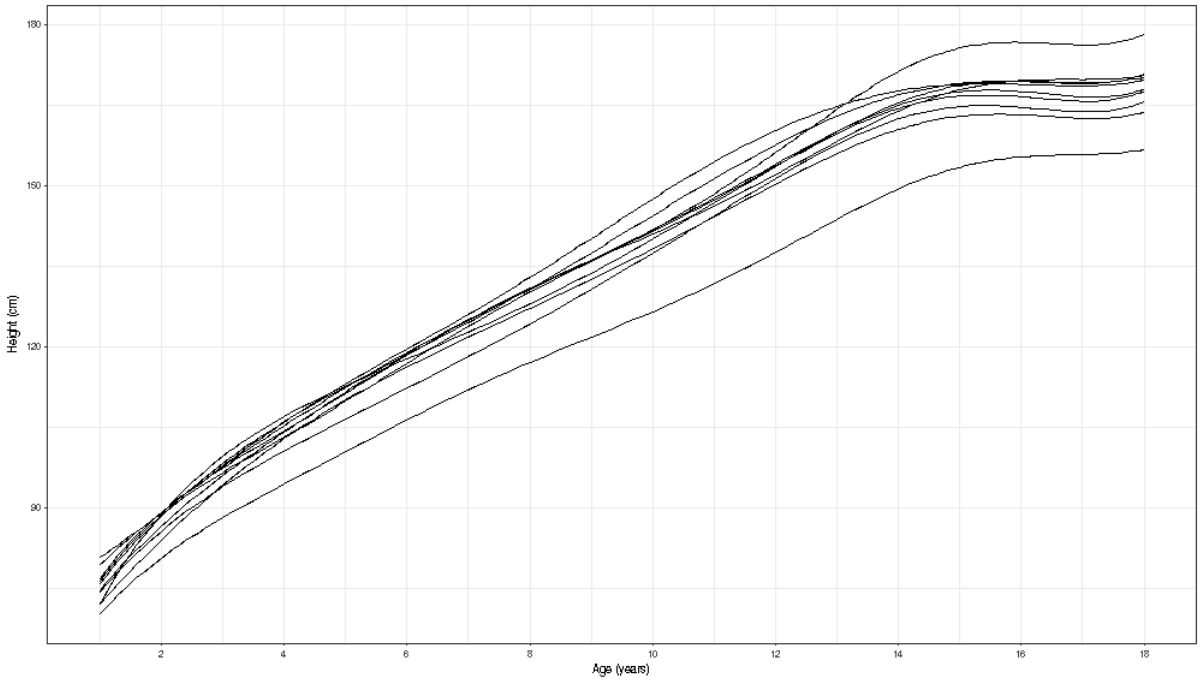

Figure 1: The heights of 10 girls measured at 31 ages, chosen randomly from the Berkley growth data set.

distribution plays a large role in any analysis. However, in a functional setting, location of an observation is no longer the sole focus of interest, but concepts such as shape, time-transformations, feature alignment, smoothness etc. play a key role. Indeed, moving to infinite dimensions also gives a rise to a range of new modes of variance in not only location, but in shape as well.

An illustrative example of such variation in shape brought about by the continuous structure of functional data can be found in the Berkley Growth Study, where the data consists of measurement records of childrens heights taken at a set of 31 ages. In Figure 1 we have illustrated 10 randomly chosen girls growth curves. Note that the measurement ages are not equally spaced. There are four measurements while the child is one year, and annual measurements from two to eight years, after which the heights are measured twice a year. These measurements reflect a smooth variation in height that could be assessed as often as desired, therefore making the data functional of nature. From a first glance the curves look very similar to one another and there doesn't appear to be much of interest in the data to be assessed. However, the features of the data are just too subtle to be seen in this type of plot and only arrive to us as the variation in the growth acceleration curves, plotted in Figure 2. Aside from the curve-to-curve variation in the growth curves, Figure 2 also serves as a good prototype for some key features one routinely encounters while working with functional data; namely variance in amplitude and phase.

With this in mind, it is not surprising that shape variation and shape outlyingness have recently received a lot of attention in literature, and there have been rigorous attempts at developing and extending functional methods that encapsulate these

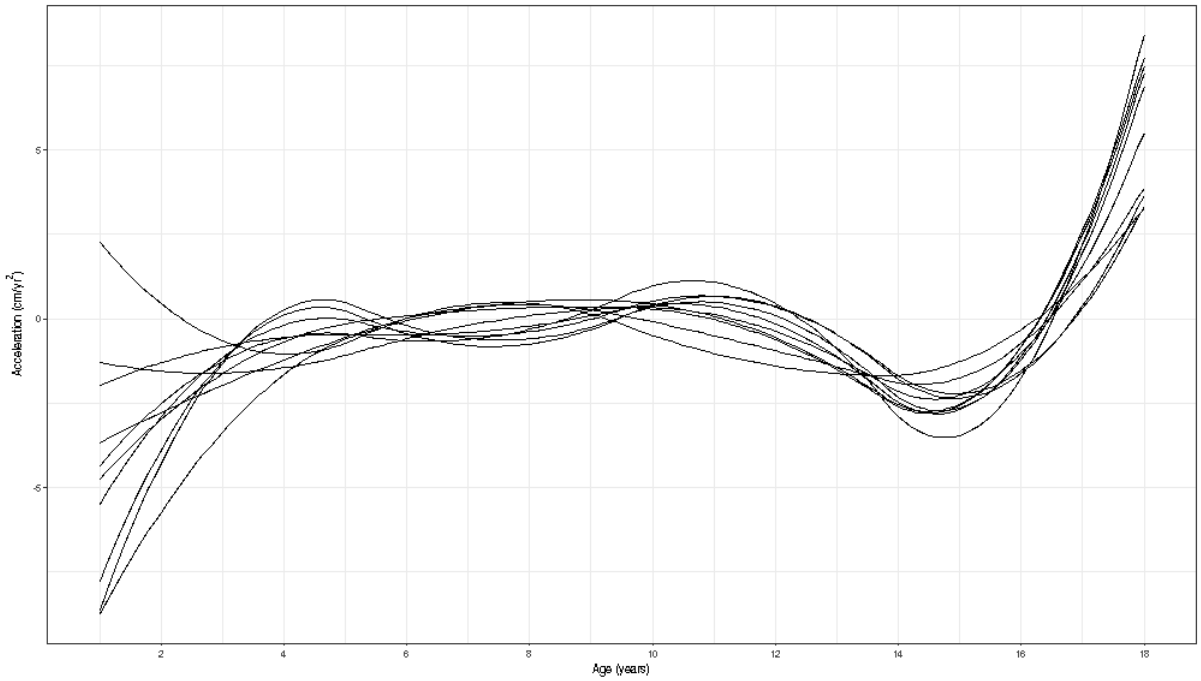

Figure 2: The estimated height acceleration curves of the 10 girls, measured in centimeters per year.

features. This has proven to be a very difficult task since the idea of encompassing shape and shape outlyingness are concepts that lack the finite dimensional basis to expand upon. In multivariate setups, ideas of outlyingness are solely interested in the location (in some sensible metric) of an observation with respect to the distribution. These simple ideas can be expanded to the the functional case as well, but the intricacies in shape and structure inherent to the functional setting still need to be addressed. However, there has been some very recent advances in the literature regarding shape and structure of functional data. Nagy et al.(2017) introduce the first formal definition of shape outlyingness, and many of the recently introduced depth functionals such as MFHD, KFSD and Pareto Depth have put great emphasis on addressing the typicality of an observation in not only location but in shape as well. See: Claeskens et al. (2014), Sguera et al. (2016) and Helander et al. (2018) for further details on these methods.

The concept of statistical depth (see for example Zuo and Serfling (2000), Tukey (1975)) was originally introduced to provide a consistent way of constructing sensible quantiles in multivariate settings where, due to the lack of natural ordering, such notions are difficult to achieve. Statistical depth provides a center-outward ordering of the data from a depth-based multivariate median outwards giving rise to regions of equal depth that have been shown to have many of the properties one would expect from quantiles (see for example: Zuo and Serfling (2000a) and Serfling (2010)). Due to these robust distributional feature revealing properties, depth has become a widely used nonparametric analytic tool. Not only does depth provide a robust measure of centrality and location, but through the depth regions it allows exploration of numerous features of the underlying distribution such as asymmetry, spread or shape (Liu et al. (1999)).

The first example of such statistical depth functional is the often referenced halfspace depth introduced by Tukey (1975),

$$HD(x, P) := \inf_{u \in \mathbb{R}^{d-1}} P[u^T(X - x) \geq 0].$$

Following the introduction of the concept, numerous other depth functionals have been defined and studied in the literature. However, it doesn't take long to recognize one glaring weakness of most of the depth notions introduced and discussed in the literature; Due to the global nature of the definition of most of these depth concepts, it is often reported that statistical depth is only suited to dealing with relatively symmetric, convexly supported unimodal distributions. Thus there have been some very recent efforts by Paindaveine and Van Bever (2013) to extend any depth method in a flexible way to be able to address these issues related to non-symmetric or multimodal distributions.

Due to the robustness and distributional feature revealing properties of multivariate depth functionals, it is not surprising that recently a lot of attention has been devoted to extending various depth notions to the functional setting. Indeed, the nonparametric nature of statistical depth makes it an attractive tool to be extended to functional setups where modeling is known to be difficult. Most of the existing approaches for functional statistical depth are solely interested in the -pointwise-

centrality of the functions, almost entirely disregarding notions of typicality in shape or structure in the distribution. However, the attention in FDA literature has shifted towards assessment of various shape properties in the data, and thus also many of the recently proposed functional depth methods have started to address typicality in shape as well (see for example Nagy et al. (2017), Helander et al. (2018) and Claeskens et al. (2014)).

Arguably shape variation and shape properties are an important aspect to consider in functional classification problems. The aim of any classification problem is to construct a rule or a metric that separates the classes as well as possible. In multivariate cases this often means that the rule is constructed such that it minimizes the within group variation, while maximizing the between group variation. The distance (in some sensible metric) of a new observation to each of the class representative cases can then be measured, and the observation is classified to belong to the group it is closest to. In functional cases this is not easy. Examples of centrally placed outliers in shape are very easy to construct (For example, see Helander et al. (2018)) which often leads to poor performance of any centrality or distance metric based attempts at separating the classes. However, functional statistical depth can perhaps provide a solution to this problem. With the recent advances in literature incorporating considerations of shape features and shape typicality in to the functional depth methods, statistical depth has become a powerful nonparametric tool for classification problems as well. This also comes with the benefit tha the usual maximum depth classification scheme is very straightforward to construct. As depth provides a measure of how -typical- an observation is within any given class (with its shape, location etc.), this information can be used to allocate the observation between the classes. In this work we consider depth based classification of functional data, especially focusing on incorporating considerations of the shape features in classification.

This thesis is organized as follows; In Section 2 we lay out the general framework for functional data establishing the usual basis on which the analysis of such data is built upon, as well as discuss statistical depth and functional depth in more detail. Section 3 introduces depth based classification schemes for functional data. In Section 4 we consider two different real data sets: the Kemijoki dataset introduced in Helander et al. (2018) and the widely used Australian weather dataset, and explore the performance of depth based classification on these data sets. Section 5 provides a short summary to the concepts introduced in this thesis as well as the results obtained.

2 General functional framework

The basic philosophy for functional data analysis is to treat the observed units as single entities arising from some continuous process, rather than as a sequence of individual discrete observations. The coined term *functional* here refers to the intrinsic continuous structure of the observed units, rather than to the explicit functional form. Indeed, this explicit functional form is unobtainable to us as due to the continuity we can never observe the functions entirely, as this would mean measuring and storing uncountably many values. Thus, by nature, functional data is only ever partially observed. In practice the observations x_i are often recorded as sequences of discrete pairs (y_{ij}, t_{ij}) where $j = 1, 2, \ldots, n_i$, y_{ij} being the snapshot of the value of the ith function at time t_{ij}. Due to this inherent partial observability of functional data, often one of the first steps in functional data analysis is to use the discrete observation sequences to reconstruct approximations of the underlying functions. This reconstruction process is discussed in detail in Section 2.1. Here we refer to t_{ij} as time as it is the most commonly encountered continuum over which the functions may be recorded, but certainly other continua such as spatial position, frequency, concentration etc. are also possible. Thus we talk about the observed units as functions, meaning that we assume the existence of a continuous function giving rise to the observed measurement sequences. By continuity we mean that in principle we could measure the function at any arbitrary point in time, as often as desired.

As each observation in a dataset is typically treated independently the same way, we shall simplify the notation by thusforth leaving out the distinction between the observations when unnecessary. Thus, we shall focus on the treatment of a single function x, observed as a sequence of pairs (y_j, t_j), $j = 1, \ldots, n$.

In addition to continuity, we often assume a certain level of smoothness from the underlying functions, so that when sampled frequently enough ('enough' depending on the level of smoothness), two adjacent data values y_j and y_{j+1} necessarily depend on one another and are unlikely to be far apart. However, this is not an inherent requirement nor the goal of FDA methods. The smoothness assumption - or the lack of it thereof - rises from our contextual knowledge of underlying process being analyzed. In some cases modeling the roughness or the noise part of a process can be where our interests lie (for example stock market pricing and rough volatility). However, in most cases we know that the underlying process or trend we are interested in is smooth and thus we might want to remove the effects of noise from our function estimates.

By *smooth* function, we often mean that the function possesses one or more derivatives, indicated by Dx, D^2x, etc. so that $D^m x$ refers to the derivative of order m and $D^m x(t)$ is the value of mth derivative at argument t. As the first step when dealing with functional data, we usually want to use the discrete data y_{ij}, $j = 1, \ldots, n_i$ to construct an estimate for the function x_i that possesses a suitable number of derivatives. As revealed by our example on the girls growth data presented in Figures 1 and 2 these various rates of change can be where the interesting variability lies.

Often however, the existence of the derivatives is not enough for us but we also want to constrain them in some way, for example to control the frequency at which the sign of the derivative can change. Without such constraints, the infinite dimensionality allows us to construct estimates that fit the observed datapoints exactly while also possessing an arbitrary number of derivatives. This is often undesirable as the raw observations can be muddled by observation error or some other kind of noise. Indeed the smoothness of the underlying function might not be apparent at all from the raw observation vector $(y_1, \ldots, y_n)$ due to the presence of noise imposed on the signal by the measurement process.

The standard way of modeling the presence of noise in the data is by an additive model where we assume that our observed sequence of values is the sum of the underlying relatively smooth function and an error term. That is, for each index j we have

$$y_j = x(t_j) + \epsilon_j,$$

where the noise or error term ϵ_j creates most of the rough variation in the raw data. The standard model for ϵ_j's is to assume many *white noise* -like properties. Namely, that they are independently distributed with mean zero and constant variance σ_ϵ^2. However, in practice many of these routinely made assumptions are violated as they are too simple for majority of functional data. For example, ϵ is often not a stationary process as the variance of the residuals itself varies over the argument t. Additionally, we can also often recognize autocorrelation in the functional residuals reflecting the fact that the rough variation brought by ϵ_j's is itself likely a result of a process with a structure we could model if needed.

Often, when the underlying function x is known to be relatively smooth, one of the tasks we may want to achieve in our function reconstruction process is to filter out the effects of this erroneous noise. The common methodologies for achieving this *smoothing* are discussed in the following Section 2.1. However, in some cases, instead of requiring smoothness from our reconstructed functions, we may choose to handle the noise by instead requiring smoothness from the results of our analysis. This is often the case with smoothed functional PCA, or smoothed functional canonical correlation analysis (Ramsay and Silverman (2005)). Although in these cases the focus is often on the informativeness of the results rather than in dealing with the noise.

While seemingly straightforward, dealing with the roughness in the raw data can be a surprisingly delicate matter. Especially when working with rapidly varying functions (functions whose first derivative changes sign at a high frequency) it is not necessarily obvious how much of the perceived roughness in the raw data is due to the noise ϵ and how much of it can be attributed to the inherent curvature of the underlying function x. Indeed, even if assumed relatively smooth, x can have strong curvature at places, often measured by the size of the second derivative as reflected in either $|D^2x(t)|$ or $[D^2x(t)]^2$. Depending on the strength of this curvature, along with the signal to noise ratio, the resolution at which the function is being observed is a key factor in determining what can be achieved through the means of FDA. Especially in the presence of noise high enough sampling frequency is essential in bringing out the features of the data; otherwise they may be lost in the noise during

the reconstrution process. Here what is -enough- depends on the level of error and curvature. When the level of error is low, a relatively low overall sampling frequency may suffice as long as the sampling is focused around areas of rapid change. As the proportion of noise in the raw data increases, a higher sampling frequency is needed. However, there are some problems related to this as well. Even if the level of noise is low, a high sampling frequency can tremendously amplify its effects if one tries to directly estimate quantities such as derivatives from the raw data, using forward differences for example. These noise related issues are often solved by fitting a linear combination of suitable basis functions to the observed datapoints. The choice of a suitable funcional basis and methods for fitting are disussed in the next Section 2.1.

2.1 From functional observations to smooth functions

Due to the inherent partial observability of functional data, the observations x often arrive to us as sequences of discrete pairs (y_j, t_j), $j = 1, \ldots, n$, where y_j is the value of the function at time t_j. Furthermore, one of the special characteritics of functional data is that the observations need not be recorded over the same sequence of measurement points t_j, but instead can each be measured over an arbitraru set of points in time, preventing the direct use of any multivariate methods in analysis. Thus, one of the first steps of any functional data analysis is to use the discrete measurement sequences to construct approximations of the underlying functional observations x. This is often done by fitting a linear combination of suitably chosen basis functions to the observed measurement sequences to represent the data as functional.

A basis function system is a set of mathematically independent functions ϕ_k with the property that given a large enough K, we can approximate any function arbitrarily well as a linear combination of K such basis functions. More formally, the goal of the function reconstruction process is to represent an observation x as a linear expansion of K known basis functions ϕ_k:

$$x(t) = c^T \phi = \sum_{k=1}^{K} c_k \phi_k(t).$$

where c is a vector of length K of the coefficients c_k, and ϕ represents a functional vector of the basis functions ϕ_k. Thus, in effect, the basis expansion represents the infinite dimensional functional observations x in terms of finite dimensional vextors c. However, this does not mean that the functional data simply reduces to multivariate data analysis; the analysis is to a great effect dependent on how the basis system ϕ is chosen.

Ideally, the chosen basis functions have characteristics that match those of the functions being estimated. This makes it possible to achieve satisfactory approximations using only a comparatively small number K of basis functions. The choice of a suitable basis system is especially important for estimating the derivatives of the observations. Ill-advised choice of the basis system may result in having to choose K large in order to attain satisfactory functional approximations, which can result in small but high-frequency oscillations in the approximations that have catastrophic

consequences in terms of derivatives. Thus, often one of the criterion for choosing a certain basis system is to also have reasonable approximations of the derivatives. Often used basis function systems are the *Fourier basis* for periodic data, and the *B-spline basis* for general functional data. In this work, we will be focusing on the B-spline basis as the more universally viable choice.

Definition 2.1 *(Fourier basis)*
Given a parameter w, the Fourier basis functions ϕ_k are given by:

$$\phi_k(t) = \begin{cases} 1, & k = 0 \\ \sin rwt, & k = 2r-1 \\ \cos rwt, & k = 2r \end{cases}$$

Definition 2.2 *(B-spline basis)*
Given a sequence of breakpoints $\tau = t_0, t_1, t_2, \ldots$, the kth B-spline $B_{k,m}$ of order m are defined recursively by

$$B_{k,1}(t) := \begin{cases} 1, & \text{if } t_i \leq t < t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

and

$$B_{k,m+1} := \frac{t - t_k}{t_{k+m} - t_k} B_{k,m}(t) + \frac{t_{k+m+1} - t}{t_{k+m+1} - t_{k+1}} B_{k+1,m}(t)$$

Then, the B-spline basis functions ϕ_k of order m with the breakpoint sequence τ are given by $\phi_k(t) = B_{k,m}(t)$.

B-splines are piecewise polynomial functions of order m, defined over a sequence of *breakpoints* $\tau = t_0, t_1, \ldots, t_L$ in a way that for splines of order $m > 1$, adjacent B-spline functions must join together smoothly at the breakpoints that separate them. Moreover, derivatives of up to order $m-2$ must also match at these junctions. The kth B-spline basis function of order m is non-zero over at most m sub-intervals, defined by the $m+1$ breakpoints $t_k, \ldots, t_{k+m+1}$. Thus, B-splines have similar computational advantages to potentially orthogonal basis systems, meaning that the computational load increases only linearly with the number of basis functions, K. As a linear combination of spline functions is itself a spline function, B-spline representations of our functional observations are smooth functions with up to $m-2$ continuous derivatives. As the order m increases, B-splines yield better approximations of both the observation x, and its derivatives.

The sequence of breakpoints τ plays a key role in defining a suitable B-spline basis. In order to gain flexibiliy in the splines we commonly increase the number of breakpoints in the region over which the functions exhibit more complex and high-frequency variation. B-splines can even model abrupt changes in the derivatives of the data by stacking breakpoints that move together. In such points, there is a loss of continuity condition for each additional coincident breakpoint, allowing us to model abrupt changes in the data at pre-determined points. This is commonly done over the edges of the interval over which the functions are observed, where in τ, the

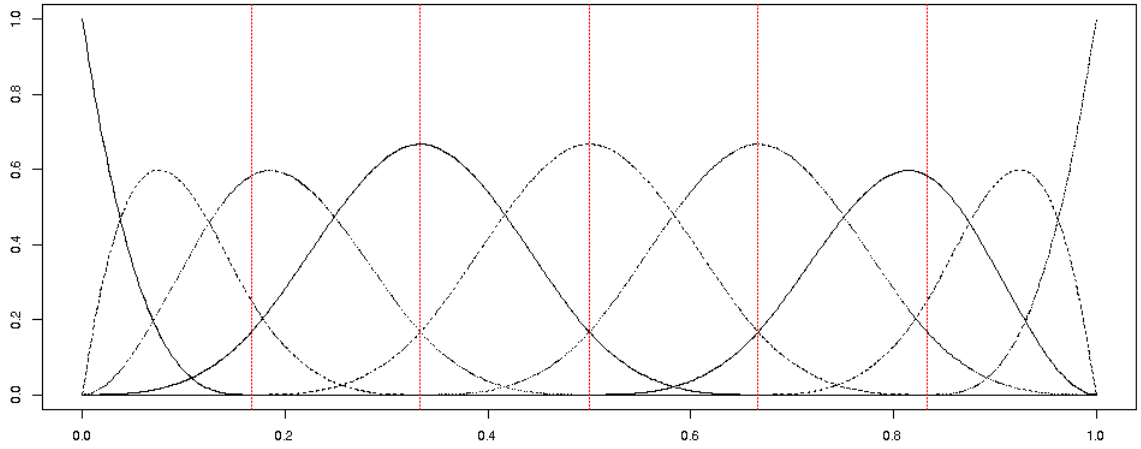

Figure 3: A cubic B-spline basis of order 4 with 5 interior breakpoints and 4 breakpoints coinciding at the edges.

first and last breakpoints t_0 and t_L appear m times, to allow modeling of open ended data. As an example, a cubic B-spline basis of order 4 with 5 interior breakpoints and 4 breakpoints coinciding at the edges is presented in Figure 3.

The common methodology for obtaining an approximation of our functional observations x is to determine the coefficients c_k of the basis function expansion by minimizing the weighted least squares criterion

$$SMSSE(y|c) = \sum_{j=1}^{n} w_j \left[y_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2,$$

where the weights w_j are determined by the covariance matrix of the residulas ϵ_j about the true curve x. Recall that the common way of modeling the presence of measurement error in the observed sequence of pairs (y_j, t_j) is by an additive model $y_j = x(t_j) + \epsilon_j$. Thus, one of the things we want to achieve in reconstructing approximations of the functional observation x is to smoothe out the effect of the often volatile measurement errors ϵ_j.

The degree of smoothing achieved is often determined by two factors. The first is the choice of the number K of basis functions used to represent the functional observations. If the number of basis functions K is equal to the number of measurement points n, we can have exact interpolation of the observed sequence, measurement error included, such that $x(t_j) = y_j$. Thus, decreasing the number K of basis functions inherently smoothes the data to a degree. Further smoothing can be achieved by the common way of penalizing the roughghness in the resulting functional approximation by adding a penalty term $PEN_m(x) = \int [D^m x(t)]^2 dt$ with respect to the derivatives of order m, often $m = 2$. This results in the commonly used roughness penalty

smoother

$$PENSSE_\lambda(y|c) = \sum_{j=1}^{n} w_j \left[y_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2 + \frac{\lambda}{n} \sum_{j=1}^{n} \left[\sum_{k=1}^{K} c_k D^2 \phi_k(t_j) \right]^2,$$

that defines a compromise between data fit and smoothness by removing the influence of the error terms ϵ_j, often seen in the form of high-frequency oscillations that can have catastrophic consequences for derivative approximations. The interested reader is encouraged to visit Ramsay and Silverman (2005) for further details of spline-smoothing and basis function representation of functional data.

2.2 Theoretical framework

In this section we lay out some important concepts often needed when working within the theoretical framework of FDA. The section is divided into two parts. First, in Section 2.2.1 we discuss the underlying *probability space* that gives rise to the distribution of functions from which we are observing our individual functions. Then, in Section 2.2.2, we recall essential definitions needed to build up towards the definition of *Hilbert spaces*, which are the spaces our observed functions live in. Finally, we conclude by discussing the theoretical functional framework adapted in this thesis, needed in the following sections. Section 2.2.1 is based on Durrett (2010), and Section 2.2.2 is based on Horváth and Kokoszka (2012).

2.2.1 Probability spaces

A *probability space* is a triple$(\Omega, \mathcal{F}, P)$, consisting of three parts; A *sample space* Ω is the set of all possible outcomes of a random process. The *σ-algebra* $\mathcal{F}$ on Ω is interpreted as a collection of events, each of which can be assigned probabilities of occuring. Finally, the *probability measure* gives a systematic way of assigning probabilities to the events in $\mathcal{F}$. The natural way of building understanding towards the definition of probability spaces begins from σ-algebras, which are in the centre of focus of probability theory.

A σ-algebra Σ on a set $\mathcal{X}$ is a collection of subsets of $\mathcal{X}$ that includes the universal set $\mathcal{X}$, and is closed under complementation and countable unions.

Definition 2.3 *Let $\mathcal{X}$ be a non-empty set and $\mathcal{P}(\mathcal{X})$ the power set of $\mathcal{X}$. Then a set $\Sigma \subseteq \mathcal{P}(\mathcal{X})$ is a σ-algebra on $\mathcal{X}$ if it satisfies the following properties:*

(i) $\mathcal{X} \in \Sigma$

(ii) if $A \in \Sigma$*, then also* $(\mathcal{X} \setminus A) \in \Sigma$*. (Closed under complementation)*

(iii) if $A_i \in \Sigma$ *for* $i = 1, 2, \ldots,$ *then also* $\bigcup_{i=1}^{\infty} A_i \in \Sigma$*. (Closed under countable unions)*

Note that due to (i) & (ii), the empty set $\emptyset$ also belongs to the σ-algebra. The primary importance of σ-algebras is the definition of a *measure*. Given a set $\mathcal{X}$ and a measure defined on $\mathcal{X}$, it would be ideal if we could measure all possible subsets

of $\mathcal{X}$. However, often this is not possible and instead we assign a measure only to some suitable subsets of $\mathcal{X}$; the collection of such subsets on $\mathcal{X}$ for which a measure is defined is necessarily a σ-algebra.

In probability theory, σ-algebra is interpreted as the collection of *events*, for which we can assign probabilities. There, the σ-algebras are often the focus of interest as the interest lies in what kinds of events the σ-algebra on a probability space contains, and on the other hand, how functions of those events (*random variables*) behave.

A non-empty set $\mathcal{X}$ together with a σ-algebra defined on $\mathcal{X}$ form a *measurable space;*

Definition 2.4 *A measurable space is a pair $(\mathcal{X}, \mathcal{A})$, where $\mathcal{X}$ is a non-empty set and $\mathcal{A}$ is a σ-algebra on $\mathcal{X}$*

Note that Measurable space doesn't need to be equipped with a measure. Instead the σ-algebra tells us which subsets of $\mathcal{X}$ will be assigned a measure.

A measure on a set $\mathcal{X}$ gives a systematic way of assigning a number, to suitable subsets of $\mathcal{X}$. Intuitively, a measure of a set is interpreted as its size, and thus acts as a generalization of the concept of volume. As such, a measure is a function that assigns a non-negative real number or ∞ to the measurable subsets (elements of the σ-algebra) on $\mathcal{X}$. For consistency, we also want a measure to be countably additive; The measure of a subset of $\mathcal{X}$ that can be decomposed to "smaller" disjoint subsets, is the sum of the measures of the "smaller" subsets. Furthermore, the empty set $\emptyset$ is given measure 0, although this is not necessarily the only subset of $\mathcal{X}$ with 0 measure.

Definition 2.5 *Let $\mathcal{X}$ be a non-empty set and Σ a σ-algebra on $\mathcal{X}$. Then a function $\mu : \Sigma \to \mathbb{R} \cup \{-\infty, \infty\}$ is a measure if it satisfies the following properties:*

(i) *For all $A \in \Sigma$, $\mu(A) \geq 0$(Non-negativity)*

(ii) *$\mu(\emptyset) = 0$ (Null empty set)*

(iii) *if $A_i \in \Sigma$ for $i = 1, 2, \ldots$ are pairwise disjoint, then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ (Countable additivity)*

Together with the properties of a σ-algebra, this results in a measure μ being monotone in the sense that if A, B measurable with $A \subseteq B$, then $\mu(A) \leq \mu(B)$.

Perhaps the most important example of a measure is the *Lebesgue measure*, which gives the usual way of assigning a measure to the subsets of $\mathbb{R}^n$. For $n = 1, 2,$ and 3, Lebesgue measure coincides with the concepts of length, area, and volume respectively. Here, as an example, we present the Lebesgue measure on $\mathbb{R}$, although it can be naturally extended to any $\mathbb{R}^n$. Let $l(I) = b - a$ denote the length of an interval $I = (a, b)$. Then, the Lebesgue outer measure $\lambda^*(E)$ of a subset $E \subseteq \mathbb{R}$ is defined as

$$\lambda^*(E) = \inf\{\sum_{k=1}^{\infty} l(I_k) : (I_k)_{k \in \mathbb{N}} \text{ is a sequence such that } E \subseteq \bigcup_{k=1}^{\infty} I_k\}$$

The Lebesgue measure is defined on the Lebesgue σ-algebra which is the collection of all sets E which satisfy the condition that, for every $A \subseteq \mathbb{R}$,

$$\lambda^*(A) = \lambda^*(A \cap E)\lambda^*(A \cap E^c).$$

For any set in the Lebesgue σ-algebra, its Lebesgue measure λ is given by its Lebesgue outer measure: $\lambda(E) = \lambda^*(E)$.

We are now ready to define a *probability space*:

Definition 2.6 *A probability space is a triple $(\Omega, \mathcal{F}, P)$ where:*

(i) The sample space Ω is an arbitrary non-empty set

(ii) The σ-algebra $\mathcal{F}$ is a set of subsets of Ω, called events

(iii) The probability measure $P : \mathcal{F} \to [0, 1]$ is a measure on $\mathcal{F}$ fulfilling the additional condition $P(\Omega) = 1$

A probability space is a type of *measure space*; That is, a measurable space $(\Omega, \mathcal{F})$, which is also equipped with a measure P. The *sample space* Ω is often interpreted as the set of all possible outcomes of a random trial. However, in our case the concept is much more abstract as no direct information of the sample space can be extracted. Instead, the underlying sample space Ω is merely thought of as some continuous set, and the focus of our attention is solely devoted to studying the σ-algebra $\mathcal{F}$ together with the *probability measure* P.

A probability measure P is a measure on the sets in $\mathcal{F}$ fulfilling the properties of Definition 2.5, with the additional constraint that $P : \mathcal{F} \to [0, 1]$ with $P(\Omega) = 1$. That is, the probability measure assigns to each set in $\mathcal{F}$, interpreted as the collection of events, a probability of the event occuring. This is also where most of our interest lies, as the probability measure P on $\mathcal{F}$ gives raise to the underlying distribution of functions from which we are observing. In practice we do not have a direct access to the events in $\mathcal{F}$ as the σ-algebra serves as an abstract mathematical tool. However, indirect information, in the form of a sample distribution P_n, can be acquired from the observed realizations of our *random function*.

In probability theory, a *random variable* is a *measurable mapping* from a probability space $(\Omega, \mathcal{F}, P)$ to a measurable space $(E, \mathcal{B})$. A measurable mapping is a function between two measurable spaces, $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, such that the preimage of any measurable set on $\mathcal{Y}$ is measurable.

Definition 2.7 *Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces, that is, $\mathcal{X}$ and $\mathcal{Y}$ are sets equipped with respective σ-algebras $\mathcal{A}$ and $\mathcal{B}$. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be measrable if the preimage of B under f is in $\mathcal{A}$, for every $B \in \mathcal{B}$.*

$$f^{-1}(B) := \{x \in \mathcal{X} : f(x) \in B\} \in \mathcal{A}, \quad \forall B \in \mathcal{B}$$

Definition 2.8 *A random variable $X : \Omega \to E$ is a measurable mapping from the probability space $(\Omega, \mathcal{F}, P)$ to a measurable space $(E, \mathcal{B})$.*

In our case, we opt to talk about a random function rather than a random variable, as the random process X we are observing maps from the underlying probability space $(\Omega, \mathcal{F}, P)$ to a *Hilbert space* $(\mathcal{H}, \mathcal{B})$ where the elements are interpreted as functions rather than vectors. Each of the functions on our Hilbert space $\mathcal{H}$ is itself deterministic. Thus the randomness is induced by the probability measure P on our underlying probability space, resulting in a distribution of functions from which our functional data arises. The next Section 2.2.2 is devoted to building understanding towards the definition of Hilbert spaces in which our observed functions lie, and ties the probability theoretical concepts laid out in this section into the FDA theory.

2.2.2 Hilbert spaces

The concept of a Hilbert space generalizes the notion of Euclidean spaces and extends the methods of vector algebra and calculus to spaces with any number of dimensions, which in our case is infinite. A Hilbert space is an abstract vector space with a structure given by an *inner product*, that is also *complete*, meaning that it contains all of its limit points allowing the techniques of calculus to be used. Note that even though we refer to the basic elements in the following definitions as *vectors*, our functions can be viewed as such vectors over an uncountably infinite index set. Thus, Hilbert spaces naturally extend many useful tools to the functional setting we are dealing with.

We begin our journey towards defining a Hilbert space by first defining an *inner product space*. Starting from the definitions of a general vector space, and an inner product;

Definition 2.9 *Let $\mathbb{K}$ be a real or complex valued scalar field. Then a $\mathbb{K}$-vector space $\mathcal{V}$ is a set of vectors $u, v, w, \cdots \in \mathcal{V}$ that is endowed with vector addition and scalar multiplication operations:*

(i) $(u, v) \mapsto u + v : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$

(ii) $(\lambda, u) \mapsto \lambda u : \mathbb{K} \times \mathcal{V} \to \mathcal{V}$

such that it adheres to the following axioms:

(i) $u + (v + w) = (u + v) + w$ *(Associativity)*

(ii) $u + v = v + u$ *(Commutativity)*

(iii) *There exists an element* $0 \in \mathcal{V}$ *such that* $v + 0 = v$ *(Identity element (addition))*

(iv) *There exists an element* $-v \in \mathcal{V}$ *such that* $v + (-v) = 0$ *(Inverse element)*

(v) $\lambda(\mu v) = (\lambda\mu)v$, *for all* $\lambda, \mu \in \mathbb{K}$ *(Compatibility)*

(vi) $1v = v$ *where 1 denotes the multiplicative identity of* $\mathbb{K}$ *(Identity element (multiplication))*

(vii) $\lambda(u + v) = \lambda u + \lambda v$, *for all* $\lambda \in \mathbb{K}$ *(Distributivity (vector))*

(viii) $(\lambda + \mu)u = \lambda u + \mu u$, *for all* $\lambda \in \mathbb{K}$ *(Distributivity (scalar))*

Definition 2.10 *Given a* $\mathbb{K}$*-vector space* $\mathcal{V}$*, an inner product is a mapping*

$$(u, v) \mapsto \langle u, v \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{K}$$

that satisfies the following properties

(i) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$

(ii) $\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$, *for all* $\lambda \in \mathbb{K}$

(iii) $\langle u, v \rangle = \langle v, u \rangle^*$

(iv) $\langle u, u \rangle \geq 0$

(v) $\langle u, u \rangle = 0 \;\Rightarrow\; u = 0$

In vector spaces, an inner product gives us refined information on not only distances, but also on the concept of *"angle"* between two elements of the space. As the concept of "angle" doesn't carry over to infinite dimensional spaces in the usual sense from finite dimensional ones, our functional inner products instead reveal information about the multiplicative amplitude of the functions over the overlapping sections of their support. This information is especially useful as it enables the concept of *orthogonality* which is a very desirable property for our functional basis system when reconstructing the functional observations. In functional context, the usual inner product is given by

$$\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R} : (X, Y) \mapsto \langle X, Y \rangle = \int_{\mathcal{V}} X(t)Y(t)\nu(dt),$$

where $\mathcal{V}$ gives the support of our Hilbert space $\mathcal{H}$ and ν denotes the Lebesgue measure. Finally, an inner product space is a general vector space equipped with an inner product.

Definition 2.11 *An inner product space is a pair* $(\mathcal{V}, \langle \cdot, \cdot \rangle)$*, where* $\mathcal{V}$ *is a* $\mathbb{K}$*-vector space and* $\langle \cdot, \cdot \rangle$ *an inner product.*

Along with the structure given by an inner product, Hilbert spaces are complete. To explore the concept of completeness, we first need a metric. Here, it is natural to only focus on metrics induced by a *norm*, since in the inner product spaces we are interested in an inner product in turn induces a norm.

Definition 2.12 *Given a* $\mathbb{K}$*-vector space* $\mathcal{V}$*, a norm is a mapping*

$$u \mapsto ||u|| : \mathcal{V} \to [0, \infty)$$

that satisfies the following properties

(i) $||u + v|| \leq ||u|| + ||v||$

(ii) $||\lambda u|| = |\lambda|||u||$, *for all* $\lambda \in \mathbb{K}$

(iii) $u \neq 0 \;\Rightarrow\; ||u|| > 0$

Then, given a $\mathbb{K}$-vector space $\mathcal{V}$ and a norm $||\cdot||$, the metric induced by $||\cdot||$ is the map

$$d : \mathcal{V} \times \mathcal{V} \to [0, \infty) : (u, v) \mapsto d(u, v) = ||u - v||.$$

Generally, a metric is a broader concept than a norm, and the metrics induced by a norm have additional properties that a general metric is not required to have. Namely, *translation invariance*: $d(u + w, v + w) = d(u, v)$ and *scaling property*: $d(\lambda u, \lambda v) = |\lambda| d(u, v)$ for any $\lambda \in \mathbb{K}$. However, in the context we are interested in, due to the following property, it is enough to focus on norm and the metric induced by a norm; On an inner product space $(\mathcal{V}, \langle \cdot, \cdot \rangle)$, the inner product induces the norm

$$||u|| := \langle u, u \rangle^{1/2}$$

called the *canonical norm* of $u \in \mathcal{V}$.

Definition 2.13 *A normed vector space is a pair* $(\mathcal{V}, ||u||)$*, where* $\mathcal{V}$ *is a* $\mathbb{K}$*-vector space and* $||u||$ *a norm on* $\mathcal{V}$*.*

Thus, every inner product space is also a normed vector space.

With normed vector spaces defined, we are now ready to work towards the definition of completeness. For this we still need to define *converging sequences* and *Cauchy sequences*.

Definition 2.14 *Let* $(\mathcal{V}, ||u||)$ *be a normed vector space. A sequence* $(u_k)_{k=1}^{\infty}$ *of vectors* $u_k \in \mathcal{V}$ *converges to* $u \in \mathcal{V}$ *if* $\lim_{k\to\infty} ||u_k - u|| = 0$*.*

Definition 2.15 *Let* $(\mathcal{V}, ||u||)$ *be a normed vector space. A sequence* $(u_k)_{k=1}^{\infty}$ *of vectors* $u_k \in \mathcal{V}$ *is a Cauchy sequence if* $\forall \epsilon > 0 \;\; \exists N_\epsilon \in \mathbb{Z}^+$ *such that* $||u_j - u_k|| < \epsilon \;\; \forall j, k \leq N_\epsilon$*.*

More informally, a Cauchy sequence is a sequence of vectors such that the distance between two consecutive elements of the sequence tends towards zero as we move along the sequence. However a Cauchy sequence might not be converging if the space $\mathcal{V}$ it lives in does not contain a suitable limit u fulfilling Definition 2.14. Thus, we say that a normed vector space is *complete*, if it contains enough of these limit points such that all of its Cauchy sequence have a limit to converge to in the space.

Definition 2.16 *A Normed space* $(\mathcal{V}, ||u||)$ *is complete if all of its Cauchy sequences converge.*

We are now finally ready to define a Hilbert space; A Hilbert space is a complete metric space with a structure given by an inner product.

Definition 2.17 *A Hilbert space* $\mathcal{H}$ *is an inner product space that is also a complete metric space.*

The functional framework often encountered in literature, and the one we will be working with here, assumes that the data points are random realizations in a Hilbert space $\mathcal{H}$. The choice of $\mathcal{H}$ is vast and depends on the type of data observed. It is often assumed that $\mathcal{H}$ is a set of functions defined over $\mathcal{V}$, a *compact* subset of $\mathbb{R}^d$, that satisfy some further regularity conditions appropriate to the context. A *compact* set is *closed*, meaning that it contains all of its limit points, and *bounded*, meaning that all of its points lie within some fixed distance of each other. Intuitively this is very natural assumption as it ensures that the functions being observed have well defined boundaries and no missing values, meaning that we can measure them at any arbitrary point in $\mathcal{V}$. In practice the dimension of $\mathcal{V}$ is often small, and for example in this thesis we will be focusing on univariate functional data.

$\mathcal{H}$ is often given additional regularity conditions to constrain the behaviour of the functions, usually to ensure that they are bounded in some suitable sense, and then equipped with an appropriate inner product. By far the most studied and assumed such space in literature is

$$\mathcal{H} = L^2(\mathcal{V}, \mathbb{R}),$$

the set of real-valued square integrable functions on $\mathcal{V}$ with respect to the Lebesgue measure ν, equipped with the inner product

$$\langle \cdot, \cdot \rangle : L^2(\mathcal{V}, \mathbb{R}) \times L^2(\mathcal{V}, \mathbb{R}) \to \mathbb{R} : (X, Y) \mapsto \langle X, Y \rangle = \int_{\mathcal{V}} X(t)Y(t)\nu(dt).$$

More formally, letting Ω denote the underlying sample space, we observe a *random function* $X = \{X(v) : v \in \mathcal{V}\} := \{X(w, v) : w \in \Omega, v \in \mathcal{V}\}$. That is, a measurable mapping $X : \Omega \to \mathcal{H}$ from the probability space $(\Omega, \mathcal{A}, P)$ to $(\mathcal{H}, \mathcal{B})$, where $\mathcal{B}$ is the σ-algebra generated by the open sets with respect to the norm induced by $\langle \cdot, \cdot \rangle$, the inner product on $\mathcal{H}$. Note that for any fixed $w \in \Omega$, $X(w, \cdot)$ is a deterministic function that maps from $\mathcal{V}$ to $\mathbb{K}$, the scalar field of $\mathcal{H}$. Thus the random function X is only random in it's first argument $w \in \Omega$, and the distribution of functions is given by P, the probability measure on the underlying probability space $(\Omega, \mathcal{A}, P)$. The distribution P is often thought to be continuous, meaning that the resulting functional space is also continuous. As an illustrative example, consider a surface $X(v, t) : [0, 1] \times [0, 1] \to \mathbb{R}^2$, and a random variable $w \in [0, 1]$ with distribution P_w. Thus, a realization of the random variable w fixes a ray along which we move on the surface, giving us the function $X(w, t) = X_w(t) : [0, 1] \to \mathbb{R}$.

Other common examples of spaces often considered in the literature include restricting $\mathcal{H}$ by adding further smoothness conditions (for example the space of continuous functions $C^k(\mathcal{V})$ on $\mathcal{V}$ with k continuous derivatives), $L^p(\mathcal{V})$ the space of Lebesgue p-integrable functions on $\mathcal{V}$ and the Sobolev space $W^{k,2}(I)$ for some appropriate k and I, a closed interval in $\mathbb{R}$. Recall that $W^{k,2}(I)$ is the space of Lebesgue square integrable functions $L^2(I, \mathbb{R})$ whose weak derivatives up to order k are also square integrable. Sometimes even more general Hilbert spaces such as multivariate functional spaces and functional manifolds are used. However, in most cases the spaces worked with in literature are assumed to possess a structure generated by an inner product. Note that the choice of the inner product is not trivial in functional spaces. In finite dimensions the choice doesn't matter as all

norms are equivalent, but in infinite dimensional cases, such as the Hilbert spaces we are working with, this equivalence doesn't hold, and the choice of the inner product affects the structure of the space.

3 Depth based classification

In this section we introduce statistical depth, as well as some of the concepts surrounding the various depth methods and their usage in statistical analysis. Starting in Section 3.1 we introduce statistical depth methods in the multivariate context where they were originally conceived. We explore some well known examples of statistical depth functions presented in literature, discuss previous work related to the properties of depth functions, and introduce the axiomatic approach formulated by Zuo and Serfling (2000) that has been widely accepted in literature. In Section 3.2, we expand our multivariate statistical depth to the functional context, and discuss the challenges brought about by the functional structure. Section 3.3 discusses the key concept of this thesis; the usage of statistical depth functions in classification problems. Finally, the following subsections are devoted to introducing the functional depth methods considered in this thesis, used in Section 4 in classification of two different real datasets, the Kemijoki dataset and the Australian weather dataset.

3.1 Statistical depth

Statistical depth functions have become a widely used nonparametric inference tool for multivariate data. Not only does statistical depth provide relevant information of centrality and outlyingness, but also reveals numerous features of the underlying distribution, such as asymmetry, spread and shape (Liu et al. 1999). Originally, the goal of statistical depth was to provide a natural center-outward ordering in multivariate context, and to give a measure of centrality for multivariate data. That is, a *depth function* $D : \mathbb{R}^d \to \mathbb{R} : x \mapsto D(x, P)$ associates to each $x \in \mathbb{R}^d$ a measure of its centrality with respect to the distribution P on $\mathbb{R}^d$. Thus, depth gives a P-based ordering for multivariate data from a depth based center outwards. As no natural ordering for multivariate setups exists, statistical depth has been suggested as the basis for defining multivariate analogues of univariate rank and order statistics, as well as to alleviate the absence of the notion of quantiles in multivariate contexts (Tukey (1975), Serfling (2010)).

Of course, to provide a meaningful center-outward ordering, a relevant notion of *center* is required. This also suggests that that points closer to that center should have higher depth, with the depth based center consisting of a set of points *globally maximizing* depth. In literature one routinely comes across three different notions of symmetry and center of symmetry, varying in generality. A standard notion of symmetry widely encountered in literature is that a random vector X in $\mathbb{R}^d$ is said to be *centrally symmetric* about θ if $(X - \theta) \sim (\theta - X)$. Liu (1990) defines X to be *angularly symmetric* about θ if $(X - \theta)/||X - \theta||$ is centrally symmetric about the origin. The broadest of these notions, introduced in Zuo and Serfling (2000) defines X to be *halfspace symmetric* about θ if $P(X \in H) \geq 1/2$ for every closed halfspace H containing θ. These notions of symmetry are presented here in an order of generality; It is well established that C-symmetry $\to$ A-symmetry $\to$ H-symmetry.

Statistical depth funtions and their properties have been profusely studied in literature. Following the introduction of the celebrated *halfspace depth* by Tukey

(1975), countless other multivariate statistical depths have been introduced. The halfspace depth HD of $x \in \mathbb{R}^d$ with respect to a distribution P is given by the minimum probability mass carried by any clsed halfspace containing x, that is,

$$HD(x, P) := \inf_{u \in S^{d-1}} P[u^T(X - x) \geq 0].$$

Liu (1990) introduced the notion of *simplical depth* SD of x, defined as the probability x belongs to a random simplex in $\mathbb{R}^d$, that is,

$$SD(x, P) := P(x \in S[X_1, \ldots, X_{d+1}]),$$

where $X_1, \ldots, X_{d+1}$ is a random sample from the distribution P, and $S[x_1, \ldots, x_{d+1}]$ denotes the d-dimensional simplex with vertices $x_1, \ldots, x_{d+1}$, the set of all points in $\mathbb{R}^d$ that are convex combinations of $x_1, \ldots, x_{d+1}$.

The *Mahalanobis depth* MhD, often attributed to Mahalanobis (1936), considered by for example Liu and Singh (1993), is defined at $x \in \mathbb{R}^d$ with respect to P as

$$MhD(x, P) := (1 + (x - \mu_P)^T \Sigma_P^{-1}(x - \mu_P))^{-1},$$

named after the *Mahalanobis distance* $d_{Mh}(x, y) = \sqrt{(x - y)^T \Sigma_P^{-1}(x - y)}$, where μ_P and Σ_P denoting the mean vector and dispersion matrix of P respectively.

Serfling (2002) introduced the *spatial depth* SSD, notable for its direct connection with the *spatial quantiles* introduced by Chaudhuri (1996), defined as

$$SSD(x, P) := 1 - ||\mathbb{E}[S(x - Y)]|| = 1 - \Big|\Big| \int S(x - y) dP(y) \Big|\Big|$$

where $||\cdot||$ denotes the usual Euclidean norm in $\mathbb{R}^d$ and $SS : \mathbb{R}^d \to \mathbb{R}^d$ is the multivariate spatial sign function given by

$$SS(x) = \begin{cases} \frac{x}{||x||}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

These and countless other examples of statistical depth functions have been considered in literature. Rousseeuw and Hubert (1999) introduced *regression depth*, Koshevoy and Mosler (1997) introduced a *zonoid depth* based on zonoid trimming, and Bartoszyński, Pearl and Lawrense (1997) introduced a depth method based on interpoint distances in the context of multivariate goodness-of-fit tests. Liu and Singh (1993) considered HD, SD and MhD along with *majority depth*, in developing methodology for assessing outlyigness of a population with respect to another, quantified by a parameter called *quality index*. Liu, Parelius and Singh (1999) considered seven different depth functions, including a depth based on outwards-in ordering via convex hull peeling, and a likelihood based depth method, and developed methodology for practical use of depth in exploratorty statistical analysis. Rousseeuw and Ruts (1996), Ruts and Rousseeuw (1996) and Rousseeuw and Struyf (1998) studied computational problems regarding depth functions and depth based contours. Zuo and Serfling (2000b) studied nonparametric notions of multivariate *scatter*

measure based on statistical depth functions. Vardi and Zhang (1999) introduced methodology for constructing depth functions based on notions of a multivariate median.

Depth functions have been vastly explored and introduced in great variety in the literature. However, this has been mostly *ad hoc* in nature as, prior to Zuo and Serfling (2000a), statistical depth didn't have a formal set of criterion to fulfill. Consequently, prior to this formal definition, there was no real systematic basis for evaluating and prefering a depth function over any other. Drawing upon the ideas discussed in Liu (1990), Zuo and Serfling (2000a) gave the first formal definition for a statistical depth function, that has since been widely embraced in the literature. The definition lists four desirable properties as criterion any statistical depth function should fulfill; *Affine invariance*, *maximality at centre*, *monotonicity relative to the deepest point*, and *vanishing at infinity*. Additionally, a depth function should be *bounded* and *nonnegative*, although, for any bounded function nonnegativity is but a matter of shift by a constant. Recall that a function f on a set $\mathcal{X}$ is bounded if there exists a $M \in \mathbb{R}$ such that for all x in $\mathcal{X}$, we have

$$|f(x)| \leq M.$$

We shall state the desirable properties first informally, and then collect them in precise notation under the Definition 3.1. For any function to effectively provide a consistent center-outward ordering of points in $\mathbb{R}^d$, it should be non-negative and bounded, and adhere to the following properties:

P1 *Affine invariance.* The depth of a point $x \in \mathbb{R}^d$ should not be dependant on the underlying coordinate system, or on the scaling of the underlying measures. Thus a statistical depth function should be invariant to any coordinate system transformations, and only consider the location of the point x relative to the distribution P.

P2 *Maximality at center.* For any distribution with a uniquely defined center (the point of symmetry with respect to some notion of symmetry), a statistical depth function should attain its maximum value at this center.

P3 *Monotonicity relative to deepest point.* As the point $x \in \mathbb{R}^d$ moves away from the depth based center along any fixed ray through the deepest point, the depth at x should decrease monotonically. Thus, intuitively, when moving away from the center of the probability mass towards the edges of the distribution we should see a corresponding monotonic decrease in depth.

P4 *Vanishing at infinity.* Intuitively the depth of a point x should approach zero as its distance from the center of the distribution grows without bounds. This property is often stated simply as the depth of a point x approaching zero as $||x||$ approaches infinity.

Collecting these into the formal definition for a statistical depth function, we have;

Definition 3.1 *(Zuo and Serfling (2000a)) Let $\mathcal{P}$ denote the class of distributions on the Borel sets of $\mathbb{R}^p$ and $P = P_X$ denote the distribution of a random vector X. The bounded and non-negative mapping $D(\cdot, \cdot) : \mathbb{R}^p \times \mathcal{P} \to \mathbb{R}$ is called a statistical depth function if it satisfies the following properties:*

P1 Affine invariance; $D(Ax+b, P_{AX+b}) = D(x, P_X)$ holds for any non singular $p \times p$ matrix A and $b \in \mathbb{R}^p$ where the $\mathbb{R}^p$ valued random vector X has distribution P_X and P_{AX+b} denotes the distribution of $AX + b$.

P2 Maximality at centre; $D(\theta, P) = \sup_{x \in \mathbb{R}^p} D(x, P)$ holds for any $P \in \mathcal{P}$ having a unique centre of symmetry θ with respect to some notion of symmetry.

P3 Monotonicity relative to the deepest point; For any $P \in \mathcal{P}$ having a deepest point θ, $D(x, P) \leq D(\theta + \alpha(x - \theta), P)$ holds for all $\alpha \in [0, 1]$.

P4 Vanishing at infinity; $D(x, P) \to 0$ as $||x|| \to \infty$, for each $P \in \mathcal{P}$.

The corresponding sample version of $D(x, P)$, denoted by $D_n(x, P_n)$ is attained by replacing P by a suitable empirical measure P_n.

Depending on the use case and the context, weaker variants of these conditions are sometimes seen. For example, due to its construction the *Pareto depth* introduced by Helander et al. (2017) is only translation invariant, and does not fulfill the propety $P4$ as stated. However, Pareto depth is intended to be used for ordering functional observations after they have been transformed to $\mathbb{R}^d$ through a collection of d measures quantifying some important features of the observed functions. Thus the geometry of the resulting dataset in $\mathbb{R}^d$ is not relevant and having translation invariance is enough for efficient ordering. For a similar argument, Pareto depth also concedes the property $P4$ in favor of a weaker variant; $P4' : D(x, P) \to 0, \text{ as } \min\{x_1, \ldots, x_d\} \to 0$.

In their work, Zuo and Serfling (2000a) also consider a significant number of existing depth functions in the light of the properties $P1 - P4$. It is reported that many of the existing depth constructions fail to satisfy one or more of these properties, thus only being effective under the specific circumstances they were introduced for. Especially *Likelihood-based* depth methods were found to generally fail to satisfy any of the $P1 - P4$, and thus are only effective under models with ellipsoidal densities, or where sensitivity to multimodality is required. Among the depth methods that were found to satisfy all of the desired properties, *halfspace depth* and *projection depth* are reported to stand favorable over their competition due to their robustness properties. These two depth methods are very similar in spirit, both being based on considering all one-dimensional projections of the dataset. This approach provides great power in extracting information, although at the cost of substantial computational load.

Due to the properties $P1 - P4$, statistical depth tends to ignore any multimodality properties of the distribution P, and is often reported to be suitable for dealing with unimodal convexly supported distributions only. To achieve sensitivity for multimodality features, the depth based center (set of points maximizing depth) should not only include the points globally maximizing depth, but the local maxima aswell. This comes with the trade off of compromising the idea of center-outward ordering, as points close to the geometric center of the distribution could be assigned

low depth. However, such multimodal or non-convexly supported distributions are met in many applications. Thus, given how versatile and powerful nonparametric modeling tool statistical depth has become, it is not surprising that there have been some recent efforts at extending depth to be able to better deal with such distributions.

Paindaveine and Van Bever (2013) provide one such extension, called *local depth*. Unlike previous attempts at local depth in the literature that typically converge to a density measure or a constant as locality becomes extreme, the construction they propose is able to take any global depth measure, and at any locality level provide a centrality measure without losing its genuine depth nature. This is achieved through first suitably symmetrizing the distribution, and then conditioning the global depth on a depth based neighbourhood, around the point of interest. To make the concept purely based on depth, the recently introduced depth based neighbourhoods from Paindaveine and Van Bever (2012) are used. The resulting local depths have interesting inferential applications and are shown to accurately capture the features of the underlying distribution, even in less well behaved cases, extending the usefulness of depth measures to a much wider range of distributions.

3.2 Functional Depth

As the measurement technology becomes ever increasingly sophisticated and the storage capacity keeps growing, we have begun to encounter more and more complex datasets commonly. This has sparked an increasing need for inferential tools for very high dimensional and even functional data. Thus, given the versatility of statistical depth in not only capturing distributional properties but also having applications in classification and modeling problems, it is not surprising that we have seen many recent attempts at extending depth to the functional context aswell. However, this is not straight forward, as direct generalizations of existing multivariate depths to functional data often neglect shape and structure properties completely, or give rise to depth constructions with absurd computational load (López-Pintado and Romo (2009), Dutta et al. (2011), Chakraborty and Chaudhuri (2014a)). For example, the commonly seen approach of integrating some pointwise centrality measure such as a multivariate depth over the domain may lead to a depth definition that misses the global and even local shape structure of the functions, focusing on the pointwise centrality only. Naive approaches to functional depth designed to fulfill the requirements for multivariate depths can even lead to degenerate depth definitions where, especially if the data is very overlapping, most if not all of the observations are given almost same depth values.

Although there have been many efforts in the literature at extending the notion of statistical depth to the functional context, functional depth still lacks a formal definition with scientific consensus. This is a problem that has resulted in a wide range of functional depth definitions with wildly differing properties and possibly only very specialized use-cases. The need for such formal definition and the lack of one thereof was first pointed out by Nieto-Reyes (2011) where a first attempt at fleshing out and addressing the problem was made. Suitable properties for functional depth

have been actively discussed in literature, and the most notable attempt at formally defining functional depth was presented by Nieto-Reyes and Battey (2016). As with the Definition 3.1 for multivariate statistical depth, we shall first informally discuss the intuition behind the proposed properties $FP1 - FP6$ below, before collecting them formally under the Definition 3.3.

FP1 *Distance invariance.* This is a property that follows in close spirit from the property $P1$ for multivariate depths. Similarly to statistical depth remaining unaffected by affine transformations in $\mathbb{R}^d$, the functional counterpart should remain unaffected by transformations through a function from $\mathcal{H}$ to $\mathcal{H}$ that (up to a scaling factor) preserve the relative distances between the elements in the d metric. This property ensures for example that depth remains unaffected by shifts such as recentering around the zero function, because, for any functional norm $||\cdot||$, $||x - y|| = ||(x - \mu) - (y - \mu)||$.

FP2 *Maximality at centre.* Statistical depth was originally introduced to provide meaning for the concept of centre of symmetry, and to act as a notion of outlyingness through giving a measure of centrality. Thus in the functional context as well, if the distribution $P \in \mathcal{P}$ possesses a unique centre of symmetry $\theta \in \mathcal{H}$ with respect to some notion of functional symmetry, the functional depth should attain its maximum at this centre.

FP3 *Strictly decreasing with respect to the deepest point.* Analogously to the property $P3$ in Definition 3.1, to achieve efficient ordering in $\mathcal{H}$, for any $P \in \mathcal{P}$ such that the deepest point z with $D(z, P) = \max_{x_{\mathcal{H}}} D(x, P)$ exists, the functional depth at a point x is required to decrease as x moves away from the depth-based centre of P. Since for some function spaces $\mathcal{H}$ there are more than one natural metric d, this requirement is formulated such that the depth D prescribes successively lower depths to functions that only lie on successively larger d-metric balls around the deepest point z. This of course also implies that $\lim_{x:d(x,z)\to\infty} D(x, P) = \inf_{x\in\mathcal{H}} D(x, P)$.

FP4 *Upper semi-continuity in x.* In $\mathbb{R}$, statistical depth and the cumulative distribution function $F_X(x) = P(X \leq x)$ are clearly linked. Indeed, depth at a point $x \in \mathbb{R}$ can be naturally defined through the cumulative distribution function, for example by $D(x, P) = \min\{P(X \leq x), P(X \geq x)\}$. Thus, in order for functional depth to retain its property of revealing features of the underlying distribution, it should be required to satisfy the same properties as a cumulative distribution function, that is, to be non-decreasing and upper-semicontinuous.

FP5 *Receptivity to convex hull width across the domain.* In practice, the functional datasets often contain subsets of the domain $L \subset \mathcal{V}$ over which the functional observations exhibit little variability, and overlap with one another significantly. Thus the property $FP5$ obligates the functional depth to give much larger weight to the values of the observations over $\mathcal{V} \setminus L$, than over L where the functions nearly coincide. This property is especially designed to negate the effects of measurement error, as over L, even the existence of relatively small

measurement error can lead to reconstructed functions that overlap in drastically different ways than when observed without measurement error, leading into drastically different ordering of the data.

FP6 *Continuity in P.* The property $FP6$ has two essential implications. For any depth function fulfilling this property, the depth based on the empirical distribution P_n converges almost surely to the population counterpart; $D(\cdot, P_n) \to D(\cdot, P)$ P-almost surely. This is an extremely important property that allows for the depth to be used in practice for statistical inference. $FP6$ also addresses the partial observability of functional data. As discussed in Sections 2 and 2.1, P_n is not accessible to us in its entirety, as the observations themselves arrive to us as discrete measurement sequences. This is a problem usually adressed through a preliminary *interpolation* or *smoothing* of the data to obtain reconstructions that approximate the functional observations. Then provided the reconstruction of the functional data is done in a way such that the empirical distribution of the *reconstructed observations*, $\hat{P}_n$ converges to the true empirical distribution, $\hat{P}_n \to P_n$ P-almost surely, $FP6$ ensures the desired convergence of the functional depth.

Collecting these in precise notation under the Definition 3.3, we have the definition for functional depth proposed by Nieto-Reyes and Battey (2016). Note that the requirement $FP5$ relies on the following preliminary definition of the convex hull of $\mathcal{H}$:

Definition 3.2 *Let $(\mathcal{H}, \mathcal{A}, P)$ be a probability space where $\mathcal{H}$ is a Hilbert space with compact support $\mathcal{V}$, $\mathcal{A}$ is the σ-algebra on $\mathcal{H}$ generated by the open d metric balls for some suitable metric d, and $P \in \mathcal{P}$, the space of all probability measures on $\mathcal{H}$. Define $\mathcal{E}$ to be the smallest set in the σ-algebra $\mathcal{A}$ such that $P(\mathcal{E}) = P(\mathcal{H})$. Then the convex hull of $\mathcal{H}$ with respect to P is defined as*

$$\mathcal{C}(\mathcal{H}, P) := \{x \in \mathcal{H} : x(v) = \alpha L(v) + (1-\alpha)U(v) : v \in \mathcal{V}, \alpha \in [0,1]\},$$

where $U := \{\sup_{x \in \mathcal{E}} x(v) : v \in \mathcal{V}\}$ and $L := \{\inf_{x \in \mathcal{E}} x(v) : v \in \mathcal{V}\}$.

Definition 3.3 *(Nieto-Reyes and Battey (2016)) Let $(\mathcal{H}, \mathcal{A}, P)$ be a probability space as in Definition 3.2. The bounded and non-negative mapping $D(\cdot, \cdot) : \mathcal{H} \times \mathcal{P} \to \mathbb{R}$ is called a statistical functional depth if it satisfies the following properties:*

FP1 Distance invariance; $D(f(x), P_{f(X)}) = D(x, P_X)$ for any $x \in \mathcal{H}$ and $f : \mathcal{H} \to \mathcal{H}$ such that for any $y \in \mathcal{H}$, $d(f(x), f(y)) = a_f \cdot d(x, y)$, with $a_f \in \mathbb{R} \setminus \{0\}$.

FP2 Maximality at centre; For any $P \in \mathcal{P}$ possessing a unique centre of symmetry $\theta \in \mathcal{H}$ with respect to some notion of functional symmetry, we have

$$D(\theta, P) = \sup_{x \in \mathcal{H}} D(x, P).$$

FP3 *Strictly decreasing with respect to the deepest point; For any* $P \in \mathcal{P}$ *such that* $D(z,P) = \max_{x\in\mathcal{H}} D(x,P)$ *exists,* $D(x,P) < D(y,P) < D(z,P)$ *holds for any* $x, y \in \mathcal{H}$ *such that* $\min\{d(y,z), d(y,x)\} > 0$ *and* $\max\{d(y,z), d(y,x)\} < d(x,z)$.

FP4 *Upper semi-continuity in x;* $D(x,P)$ *is upper semi-continuous as a function of* x, *i.e., for all* $x \in \mathcal{H}$ *and all* $\epsilon > 0$ *there exists a* $\delta > 0$ *such that*

$$\sup_{y:d(x,y)<\delta} D(y,P) \leq D(x,P) + \epsilon.$$

FP5 *Receptivity to convex hull width across the domain;* $D(x,P_X) < D(f(x), P_{f(X)})$ *for any* $x \in \mathcal{C}(\mathcal{H},P)$ *as in Definition 3.2 with* $D(x,P) < \sup_{y\in\mathcal{H}} D(y,P)$ *and* $f : \mathcal{H} \to \mathcal{H}$ *such that* $f(y(v)) = \alpha(v)y(v)$ *with* $\alpha(v) \in (0,1)$ *for all* $v \in L_\delta$ *and* $\alpha(v) = 1$ *for all* $v \in L_\delta^C$, *where:*

$$L_\delta := \arg\sup_{H\subseteq\mathcal{V}} \left\{ \sup_{x,y\in\mathcal{C}(\mathcal{H},P)} d(x(H), y(H)) \leq 0 \right\}$$

for any $\delta \in [\inf_{v\in\mathcal{V}} d(L(v), U(v)), d(L,U))$ *such that* $\lambda(L_\delta) > 0$ *and* $\lambda(L_\delta^C) > 0$.

FP6 *Continuity in P; For all* $x \in \mathcal{H}$, *for all* $P \in \mathcal{P}$ *and for every* $\epsilon > 0$, *there exists a* $\delta(\epsilon) > 0$ *such that* $|D(x,Q) - D(x,P)| < \epsilon$ *P-almost surely for all* $Q \in \mathcal{P}$ *with* $d_P(Q,P) < \delta$ *P-almost surely, where* d_P *metricises the topology of weak convergence.*

Unlike in the multivariate case, for functional data its far more difficult to summarize which properties a functional depth should satisfy. Although the contribution of Nieto-Reyes and Battey (2016) towards formally defining functional depth, and the survey of the properties of various such depths existing in the literature is highly relevant, the ideas presented in the paper have not reached similar consensus to the widely accepted Definition 3.1 for the multivariate counterpart. Even though the properties $FP1 - FP6$ and the intuition behind them makes sense for a depth function, they have received some critique in the literature and might need further exploration or reformulation (Gijbels and Nagy (2017)).

In functional spaces, the property $FP1$ was found to be very demanding without further restrictions, and a depth definition fulfilling this can be hard to achieve. Yet, suitable invariance properties are desirable to allow the use of depth as a tool for comparisons between distributions. Therefore, in literature, the invariance of depth with respect to more specific types of mappings, specifically, *function-affine mappings* and *scalar-affine mappings* is often considered. The condition $FP2$ is a straightforward translation of $P2$. However, even in the multivariate context there exists no unique notion of symmetry. Therefore, Nieto-Reyes and Battey (2016) considered the alternative property, $FP2G$: *Maximality of D at a Gaussian process mean*, which is a straightforward translation of property $P2$ towards functional data, stating:

FP2G *Maximality at Gaussian process mean:* For P a zero-mean, stationary and almost surely continuous Gaussian process on $\mathcal{V}$, $D(\theta,P) = \sup_{x\in\mathcal{H}} D(x,P) \neq \inf_{x\in\mathcal{H}} D(x,P)$, where θ is the zero mean function.

Finally, the property $FP5$ might not be suitable for a -general- definition of functional depth. Although the property is sensible in some applications where measurement error is of special concern, it is too restrictive in general, and in some cases can lead into excluding information that would give valuable insight into the nature of the process being analyzed. Furthermore, $FP5$ has been found to have negative implications on the, as a property, more desirable invariance with respect to function-affine transformations, often considered in literature.

Typically, the functional depth approaches encountered in the literature can be roughly divided into the following two classes. The first style of approach integrates some centrality measure, often a suitable multivariate depth, over the domain of the observations. This category includes for example approaches such as the integrated depth (Fraiman and Muniz (2001)), the random projection depth and the double random projection methods discussed in (Cuevas et al. (2007)), the integrated dual depth (Cuevas and Fraiman (2009)), the (modified) band depth (López-Pintado and Romo (2009)), the (modified) half region depth (López-Pintado and Romo (2011)), and the multivariate functional halfspace depth (Claeskens et al. (2014)). The second class of definitions consists of notions that aim to provide an expected distance from the function x to the distribution of functions P. This class includes approaches such as the h-mode depth (Cuevas et al. (2007)), the functional version of spatial depth (Chakraborty and Chaudhuri (2014a,b)), and the recently introduced kernelized version of functional spatial depth (Sguera et al. (2016)).

Most approaches presented in the literature focus on the pointwise centrality of the functions as a measure of depth in the distribution P. As a result they ignore important shape related features unique to functional data. This often leads to centrally placed shape-outliers being given high depth values. This may result in poor or unpredictable behaviour in applications such as classification - one of depth's primary use-cases in the functional context.

Examples of central shape outliers that can even be centrally placed in one or more derivatives, yet be clearly outlying in shape, are not difficult to construct. For example, let A, B be random variables from distributions F_A and F_B with expected values μ_{F_A} and μ_{F_B} respectively. Consider $\mathcal{F}$, a set of n observations each of the form $f_i(x) = A_i x + B_i$, with an outlier function $o(x) = \hat{\mu}_{F_A} x + C \sin(wx) + \hat{\mu}_{F_B}$, where $\hat{\mu}_{F_A}$ and $\hat{\mu}_{F_B}$ are the sample means of the observed distributions of A and B, $\hat{F}_A$ and $\hat{F}_B$ respectively, C is a scaling parameter dependent on the spread of $\hat{F}_B$, and w is a frequency parameter. Then, with a suitable choice of C, $o(x)$ is clearly central in $\mathcal{F}$, as well as its derivative function $o'(x) = \hat{\mu}_{F_A} + C\cos(wx)$ being central in $\mathcal{F}'$, the set of the derivative functions of $\mathcal{F}$. To illustrate this example, a random sample of 10 such curves $f(x)$ was simulated. The parameters A and B were chosen randomly from the uniform distributions on the intervals $[0.5, 1.5]$ and $[-1, 1]$ respectively. Then an outlying curve o was added with the parameters $C = 0.15$ and $w = 10$. The random curves together with their derivatives are presented in Figure 4, with the outlying curve o highlighted in black. Of course, this is a very simple pathological example, and the interested reader is encouraged to visit for example Cuevas et al. (2007), Claeskens et al. (2014), Nagy et al. (2017) or Helander et al. (2017) for more thorough examples and discussion.

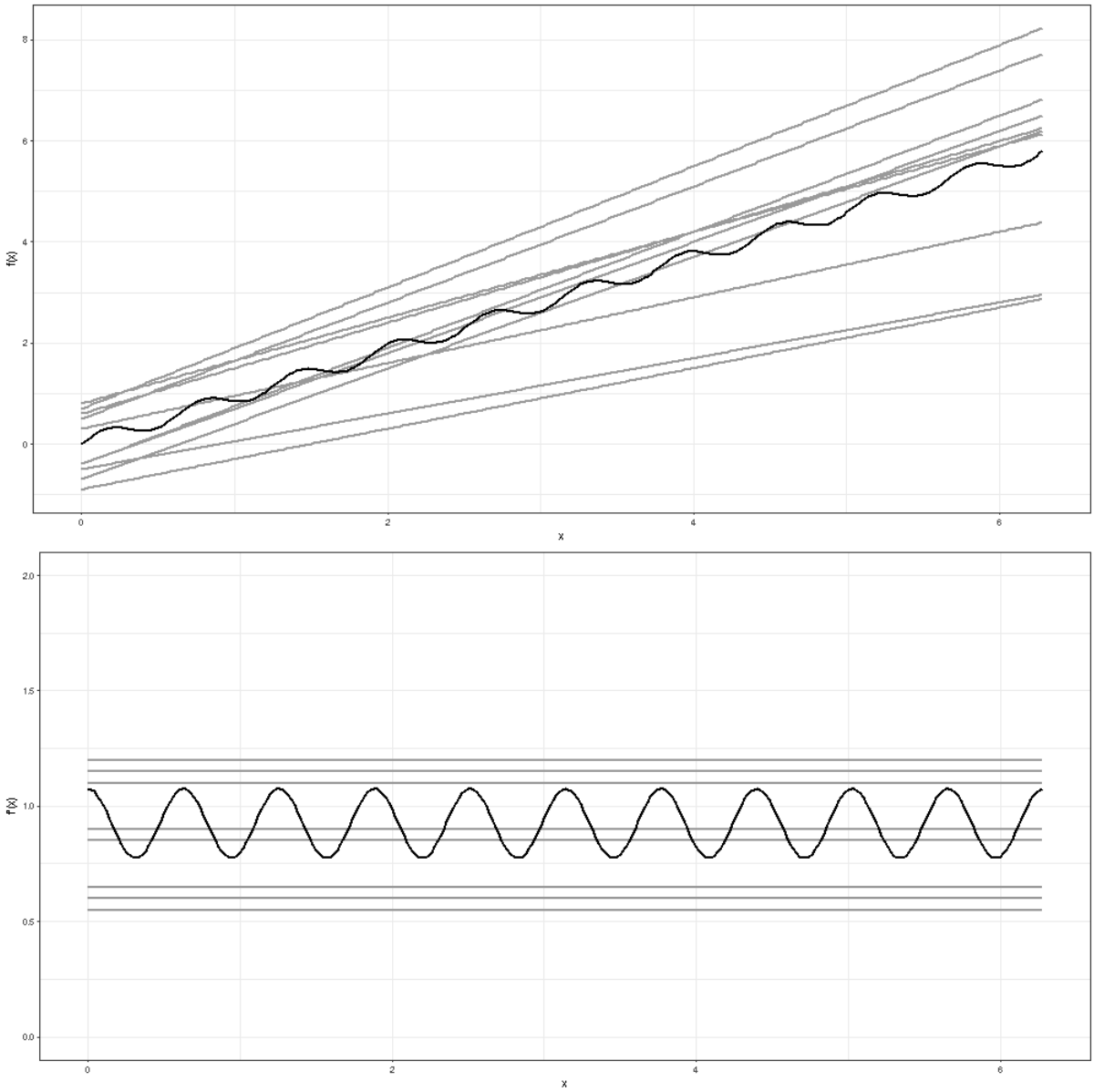

Figure 4: Example of a centrally located shape-outlyer which is also a centrally located shape-outlyer in its first derivative.

Shape and shape outlyingness for functional data have recently received a lot of attention in the literature. Especially the focus of the discussion has revolved around identifying observations differing or outlying merely in shape, a task functional depth has been reported ineffective at. As a partial remedy to the problem it has been proposed for the functional depth method to, alongside with the observed curves, take into consideration derivatives of various degrees or some other auxiliary curves such as time registration curves. This approach is mainly applied in the case of integrated depths (Fraiman and Muniz (2001)) and infimal depths (introduced by Mosler (2013)) where, instead of considering a pointwise univariate depth over the

domain, we consider a multivariate depth over the pointwise multivariate sample. More formally, let $\mathcal{H}$ be a Hilbert space with a compact support $\mathcal{V}$ and $P \in \mathcal{P}$, the space of all probability measures on $\mathcal{H}$. Consider $x \in \mathcal{H}$, and let $x(t)$ denote the value of the function x at t, and $P_{X(t)}$ denote the corresponding marginal distributions of $X \sim P$, for all $t \in \mathcal{V}$. Then, the classical way of defining integrated and infimal depths, with respect to the univariate depths $D(x(t), P_{X(t)})$, are

$$FD(x, P) = \int_{t \in \mathcal{V}} D(x(t), P_{X(t)}) dt$$

and

$$ID(x, P) = \inf_{t \in \mathcal{V}} D(x(t), P_{X(t)})$$

respectively. Then, to incorporate consideration for shape into the depth definition, one might instead consider for example the functions x together with their derivatives x', leading into the following two depth functionals

$$FD^{(2)}(x, P) = \int_{t \in \mathcal{V}} D((x(t), x'(t))^T, P_{(X(t),X'(t))^T}) dt$$

and

$$ID^{(2)}(x, P) = \inf_{t \in \mathcal{V}} D((x(t), x'(t))^T, P_{(X(t),X'(t))^T}),$$

where $D((x(t), x'(t))^T, P_{(X(t),X'(t))^T})$ denotes the multivariate depths of the pointwise multivariate sample $(x(t), x'(t))^T$ of the function values at $t \in \mathcal{V}$, with respect to $P_{(X(t),X'(t))^T}$, the joint marginal distribution of X and X' at $t \in \mathcal{V}$. Note that, instead of considering only the the derivative functions x' (of order one or more) as in the example, other auxiliary curves, such as for example registration curves, containing information of the features of x can be considered. Recall that registration curves $\phi_i : \mathcal{V} \to \mathcal{V} : t \mapsto \phi_i(t)$ are monotonically increasing and continuous transformations of $t \in \mathcal{V}$, constructed such that the key features of x_i align for the composed functions $(x_i \circ \phi_i)(t) = x_i(\phi_i(t))$.

However, as the approach is often used in conjunction with derivatives of various orders, it relies on x and the random function X both being (almost surely) differentiable. Thus, the outcome of the method depends crucially on not only the method chosen to estimate these derivatives, but also on the way the functional observations are reconstructed in practice. This plays an enormous role when considering derivatives of higher order, especially in the presence of measurement error. Furthermore, as the approach proceeds with pointwise consideration of the multivariate depths, it misses the global features of the data.

Another approach discussed in literature for incorporating consideration for shape in functional depth is to map the original functions to a different functional or multivariate space through some suitable projections or transformations, and then to proceed with known functional or multivariate depth methods. This approach was proposed for example by Helander et al. (2018). The method adopted in the paper first maps the functional observations into a multivariate space through a set of measures called *statistics of interest (SOI)* that quantify some important

and interesting features of the data, and then proceedes to assign depths to the observations based on a new multivariate depth definition, *Pareto Depth*, applied on the vector of the SOI. The approach was demonstrated to achieve very powerful result in supervised classification, as the flexible choice of SOI allows measuring and focusing on the important features of the data enabling accurate distinction between the classes. However, while powerful, the method is constructed ad hoc and heavily relies on the analysts contextual knowledge of the data and the phenomenon at hand.

Despite the recent attention shape has received in the FDA context, previously, part of the difficulty has been due to the lack of an accurate definition describing shape outlyingness. However, such definition was given recently by Nagy et al. (2017), who considered depth-based recognition of shape outlying functions and introduced definitions and methodology for enabling functional depth better receptivity for differences in shape. Assume that some particular definition for outlyingness in the multivariate context is agreed upon. For example, define a multivariate outlier as an observation with a particularly low depth with respect to the rest of the random sample. Then, a functional observation can be classified as a functional outlier through the following recursive definition:

Definition 3.4 *(Nagy et al. (2017)) Let be $\mathcal{H}$ a Hilbert space with a compact support $\mathcal{V}$. Let $P \in \mathcal{P}$, the space of all probability measures on $\mathcal{H}$, $X \sim P$ and $x \in \mathcal{H}$. If there exists $t \in \mathcal{V}$ such that $x(t) \in \mathbb{R}$ is outlying with respect to $P_{X(t)}$, the marginal distribution of X at t, then we say that x is a 1st order outlier with respect to P.*

For $J = 2, 3, \ldots$, assume that the collections of jth order outliers with respect to P are given for $j = 1, \ldots, J-1$. If there exists a set of points $(t_1, \ldots, t_J)^T \in \mathcal{V}^J$ such that $(x(t_1), \ldots, x(t_J))^T \in \mathbb{R}^J$ is outlying with respect to $P_{(X(t_1),\ldots,X(t_J))^T}$, the joint marginal distribution of X at points $t_1, \ldots, t_J$, and at the same time x is not a jth order outlier with respect to P for any $j = 1, \ldots, J-1$, then we say that x is a Jth order outlier with respect to P.

According to this definition, 1st order outliers are functions outlying in *location* usually considered as the outliers in literature. The higher order outliers for $J = 2, 3, \ldots$ on the other hand would be functions outlying in shape. As the joint distribution $P_{(X(t_1),X(t_2))^T}$ relates to the difference of functional values $X(t_1)$ and $X(t_2)$, the 2nd order outliers are the curves violating the pattern of the functions from P in terms of linear growth. Similarly, 3rd order outliers violate the pattern in convexity / concavity (acceleration) etc. Thus, in essence, 2-dimensional projections can be used to emulate differences in first derivative, 3-dimensional projections to emulate differences in 2nd derivative and so forth.

Based on these observations and the idea of the outlyingness from Definition 3.4, Nagy et al. (2017) presented the following definitions for Jth order integrated and infimal depths, that respectively quantify the *average* and *maximum* outlyingness of an observation, up to order J:

Definition 3.5 *For $J = 1, 2, \ldots$ the Jth order integrated depth of $x \in \mathcal{H}$ with respect to $X \sim P \in \mathcal{P}$ is defined as*

$$FD^J(x, P) := \int_{\mathcal{V}} \cdots \int_{\mathcal{V}} D((x(t_1), \ldots, x(t_J))^T, P_{(X(t_1),\ldots,X(t_J))^T}) dt_J \ldots dt_1.$$

Moreover, the Jth order infimal depth of x with respect to $X \sim P$ is defined as

$$ID^J(x, P) := \inf_{t_1,\dots,t_J \in \mathcal{V}} D((x(t_1), \dots, x(t_J))^T, P_{(X(t_1),\dots,X(t_J))^T}).$$

Note that due to the invariance properties of statistical depth, in the above definition, D only needs to be evaluated for points $(t_1, \dots, t_J)^T \in \mathcal{V}^J$ such that $t_J \leq \cdots \leq t_1$. Furthermore, Nagy et al. (2017) showed a direct connection between the above defined Jth order depth functions and the depth of x at point t considering the derivatives of x up to order J, assuming x is J-times differentiable. Later, Nagy et al. (2018) further expanded upon the Definition 3.5 providing the following definition of the Jth order integrated moment depth and exploring its properties;

Definition 3.6 *Let $J \in \mathbb{N}$ and $k \geq 1$. For $x \in \mathcal{H}$, a Hilbert space with compact support $\mathcal{V}$ and $P \in \mathcal{P}$, the space of all probability measures on $\mathcal{H}$ the Jth order kth moment integrated depth of x with respect to P is given by*

$$FD_k^J(x, P) := \left(\int_{\mathcal{V}} \cdots \int_{\mathcal{V}} \left| D((x(t_1), \dots, x(t_J))^T, P_{(X(t_1),\dots,X(t_J))^T}) + \frac{1}{2} \right|^k dt_J \dots dt_1 \right)^{1/k} - \frac{1}{2},$$

where the depth D stands for the usual multivariate halfspace depth in $\mathbb{R}^J$.

Due to the substantial computational burden the multiple integrals impose, in practice it is not feasible to compute the FD_k^J in full. Thus, for $M \geq 1$ consider a random sample $(T_{1,1}, \dots, T_{1,J})^T, \dots, (T_{M,1}, \dots, T_{M,J})^T \in \mathcal{V}^J$ of M combinations of J time indices from the uniform distribution on $\mathcal{V}^J$, with no permutations. Then, the approximated version of FD_k^J is given by

$$AFD_k^J(x, P_n) = \left(\frac{1}{M} \sum_{m=1}^{M} \left| D((x(T_{m,1}), \dots, x(T_{m,J}))^T, P_{(X(T_{m,1}),\dots,X(T_{m,J}))^T}) + \frac{1}{2} \right|^k \right)^{1/k} - \frac{1}{2}.$$

In most cases the approximated version of the Jth order integrated moment depth provides reliable results if M is chosen large enough. Nagy et al. (2017) explored similar treatment for the Jth order integrated depth of Definition 3.5 and suggested the following approach for choosing M; First, the depths of each datapoint is approximated $m* > 1$ times independently, using different choice of the parameter $M = M_j$ for each replication. Then, the value of M is chosen as the smallest M_j such that the average correlation coefficient of the $m*$ vectors of depths corresponding to M_j exceeds some chosen treshold value, for example 0.99. Thus, the final choice of M has little effect on the resulting ordering of the curves without sacrificing computational efficiency. In this thesis, values of M between 5000 and 10 000 were chosen.

Following in the spirit of Nagy et al. (2018), in Section 4 we shall focus on the application of AFD_k^J in supervised classification problems for two different real datasets, comparing its performance to two other recently proposed depth methods, the multivariate functional half-space depth (MFHD) (Claeskens et al. (2014)) and the kernelized functional spatial depth (Sguera et al. (2016)).

3.3 Depth based classification

The natural theoretical framework of any supervised classification problem is given by the random pair (Y, G), where Y is a multivariate or functional random variable and G is a categorical random variable expressing the class membership. Here, we focus on functional classification problems and as such, Y is a functional random variable. Often for each class $G = g$, we have $Y = Y_g \sim P_g$ where P_g denotes the distribution associated the gth class. From now on, we assume that G only takes values 0 or 1 and that we observe the sample $(y_i, g_i),\ i = 1, \ldots, n$ of $n = n_0 + n_1$ pairs taken from the distribution of (Y, G), such that we have n_0 observations from the group 0 and n_1 observations from the group 1. Additionally, we have an independent random curve x distributed as Y but with unknown class membership G_x. Thus, using the information contained in the observed sample of the pairs (y_i, g_i) the goal of the supervised functional classification problem is to provide a rule that predicts G_x.

There is a rich body of work surrounding supervised functional classification and several such methods have been proposed in the literature. For example, Marx and Eilers (1999) considered the application of generalized linear regression model to functional supervized classification using a P-spline approach. James and Hastie (2001) proposed a functional version of the multivariate linear discriminant analysis, applied on spline reconstructions of the observations. Hall et al. (2001) suggested a dimension reduction approach using functional principal component projections and solving the resulting multivariate problem with discriminant analysis or kernel methods. Ferraty and Vieu (2003) developed functional classifiers based on kernel estimators of prior probabilities. Biau et al. (2005) and Cérou and Guyader (2006) considered the extension of k-nearest neighbors method and its properties in infinite dimensional spaces. Epifano (2008) developed classifiers based on functional shape descriptors. Finally, Delaigle and Hall (2012) considered classifiers based on dimension reduction through partial least squares or carefully chosen functional principal component projections, and studied their asymptotic properties.

Due to its many desirable properties, depth based methods have also been considered for functional supervised classification problems. The main difference between the depth based classification procedures and the ones mentioned above is that, due to depth being a measure of typicality and outlyingness, the depth based methods are specifically designed to be suitable for datasets that may contain outlying curves. There are three examples of depth based supervised functional classification methods in the literature; *Distance to the trimmed mean* procedure and *weighted average distance* procedure, that were considered by López-Pintado and Romo (2006), and *within maximum depth* procedure first considered in the multivariate context by Ghosh and Chaudhuri (2005), and then for functional data by Cuevas et al. (2007).

Given a proportion α, the *distance to the trimmed mean* procedure computes the mean of the $1 - \alpha$ deepest curves of each group m_g^α, called the α-trimmed mean, and assigns x to the group which minimizes $||x - m_g^\alpha||$. Thus, basing the estimation of the mean on the chosen proportion of the deepest, -most typical-, functions allows to

obtain robust mean functions that more accurately describe the different classes. In the *weighted averaged distances* procedure, for each group, the weighted averages of the distances $||x - y_i||,\ i = 1, \ldots, n_g$ are computed such that the weights are given by the within-group depth values of y_i, $D(y_i, \hat{P}_g)$, where $\hat{P}_g$ denotes the empirical distribution of the class g. Then, x is assigned to the group for which the weighted averaged distances are minimized. Finally, the *within maximum depth* procedure computes the depth value of x with respect to the empirical distributions of each group, and assigns x to the group in which the highest depth is achieved.

Generally, any functional depth can be used in conjunction with any of the depth based methods discussed above to perform supervised functional classification. However, given the discussion in Section 3.2 on the importance of shape for functional data, in order to ensure the classification accurately captures the features of the underlying functional distribution, depths receptive to the shape properties should be considered. Thus, in Section 4 we will apply the within maximum depth classification procedure to two different real datasets using the approximated version of the Jth order integrated moment depth (AFD_k^J), the multivariate functional halfspace depth (MFHD) (Claeskens et al. (2014)) and the kernelized functional spatial depth (KFSD) (Sguera et al. (2016)), and compare their performance in supervised functional classification.

4 Real data examples

In this section, we consider two different real data examples, the Kemijoki dataset and the Australian rainfall dataset. For both datasets, we consider the performance of three recently introduced functional depth definitions in supervised functional classification, using the within maximum depth classification procedure as described in Section 3.3. The functional depths considered are:

(i) the approximated version of the Jth order kth moment integrated depth AFD_k^J (Nagy et al. (2018)) for different values of J and k,

(ii) the multivariate functional halfspace depth (MFHD) (Claeskens et al. (2014)) applied to the functional observations and their derivatives, and

(iii) the kernelized functional spatial depth (KFSD) (Sguera et al. (2016)) with a Gaussian and automatic bandwidth selection provided by the authors.

A leave-one-out classification scheme was conducted between each pairing of classes within each dataset. One at a time, for each classification pairing, each observation was taken out of the pooled sample and then classified to the group with respect to which it had the higher depth value. In case of ties, the group belonging was chosen randomly, weighted by the relative sizes of the two groups. The classification was conducted for each pairing in both datasets, using each of the three depth functions discussed above.

Additionally, for comparison, a similar leave-one-out classification scheme was conducted between the same pairings using the principle component (PC) and partial least squares (PLS) classifiers introduced by Delaigle and Hall (2012). The methods are based on calculating a statistic of the form $T(x) = (\langle x, \varphi\rangle - \langle \hat{\mu}_1, \varphi\rangle)^2 - (\langle x, \varphi\rangle - \langle \hat{\mu}_0, \varphi\rangle)^2$ for each observation x, where $\langle x, \varphi\rangle = \int_{\mathcal{V}} x\varphi$ and φ is a function on $\mathcal{V}$. x is then classified based on the sign of the statistic to belong to group 1 when $T(x)$ is positive or to group 0 when $T(x)$ is negative. The classifier is based on the careful selection of the function φ to guarantee optimal classification properties (see Delaigle and Hall (2012) for more details). Both methods for choosing φ suggested by the authors, PC and PLS, were explored.

Detailed descriptions of both datasets and the pairwise classification problems, as well as the leave-one-out misclassification rates, are presented in Sections 4.2 and 4.1.

4.1 Australian weather dataset

The Australian rainfall dataset[1] (first analysed by Lavery et al. (1992)) depicts the daily rainfall measurements at the 191 Australian weather stations, taken over the interval from January 1840 to December 1990. Each of observed functions depicts the rainfall measurements over a year at one of the 191 weather stations. For each station, the rainfall measurement at time t is given as an average over those years for which the station had been operating, where this averaging is given by a local polynomial

[1] Available at https://rda.ucar.edu/datasets/ds482.1/

smoother passed through the discrete observations. The 190th station was left out, as its rainfall pattern was clearly outlying in shape compared to those of all the other stations. The data is presented in Figure 5 with randomly chosen observations highlighted in solid black line to help perceive the overlapping observations.

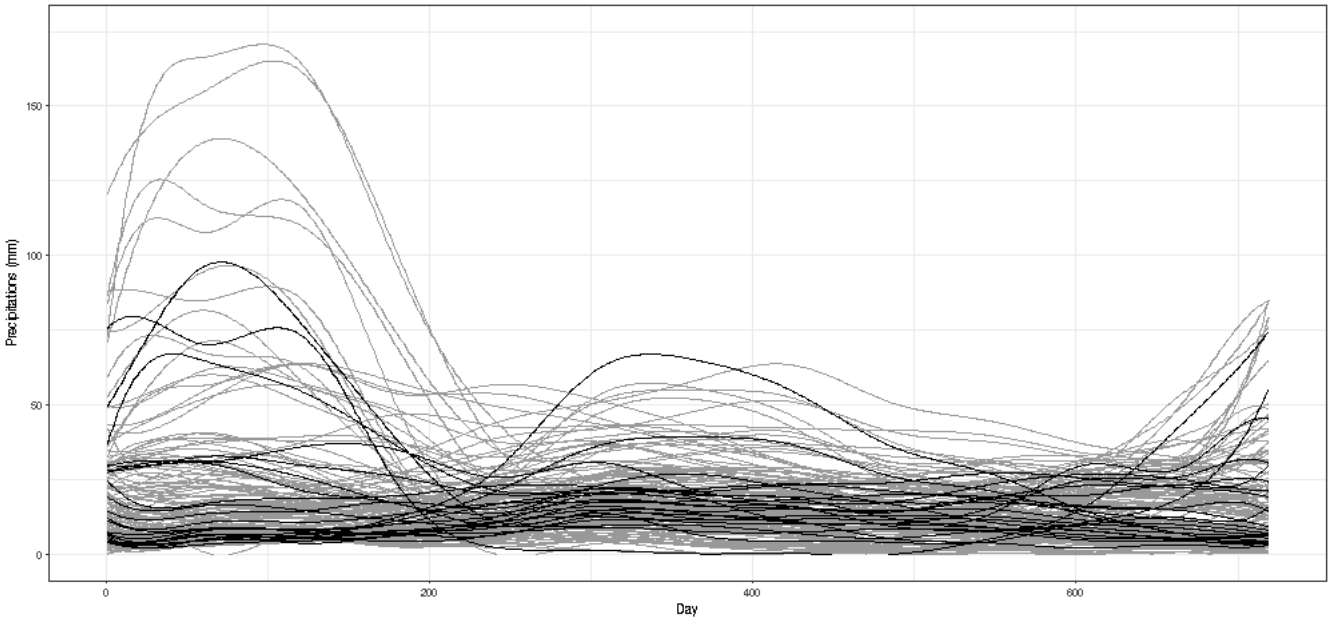

Figure 5: Yearly Australian rainfall data from the 190 weather stations with randomly chosen observations highlighted in black.

There are two natural clusterings of the data based on the geographical location of the stations, for both of which the shape properties between the clusters are much more distinguished than for the Kemijoki dataset. The commonly considered clustering, presented in Figure 6 divides the observations to North (top) and South (bottom) clusters. This clustering is also very natural from the perspective of the rainfall pattern, as it divides the observations to a group with a "tropical" pattern with most of their rainfall over the year falling on the summer months (north cluster), and to a group with most of its rain on the cooler months (south cluster). The three observations that obtained the highest depth values for AFD_1^2 in each cluster are highlighted in Figure 7. Along with the North-South (NS) clustering, the West-East (WE) clustering, presented in Figure 8, was considered. Although not quite as distinct, the WE clustering provides a clear separation in shape pattern aswell. The West cluster (top) typically has a more distinguished amplitude variance over the year, with sharply defined seasons where the precipitation falls on either warm or cold months. However, for the East cluster (bottom), the rain falls much more uniformly over the year with little to no separation between the seasons, and a stable precipitation throughout the year instead. The three observations that obtained the highest depth values for AFD_1^2 in each cluster are highlighted in Figure 9.

The geographical location of each of the weather stations based on which the two clusterings were decided is presented in Figure 10. The North (grey) to South

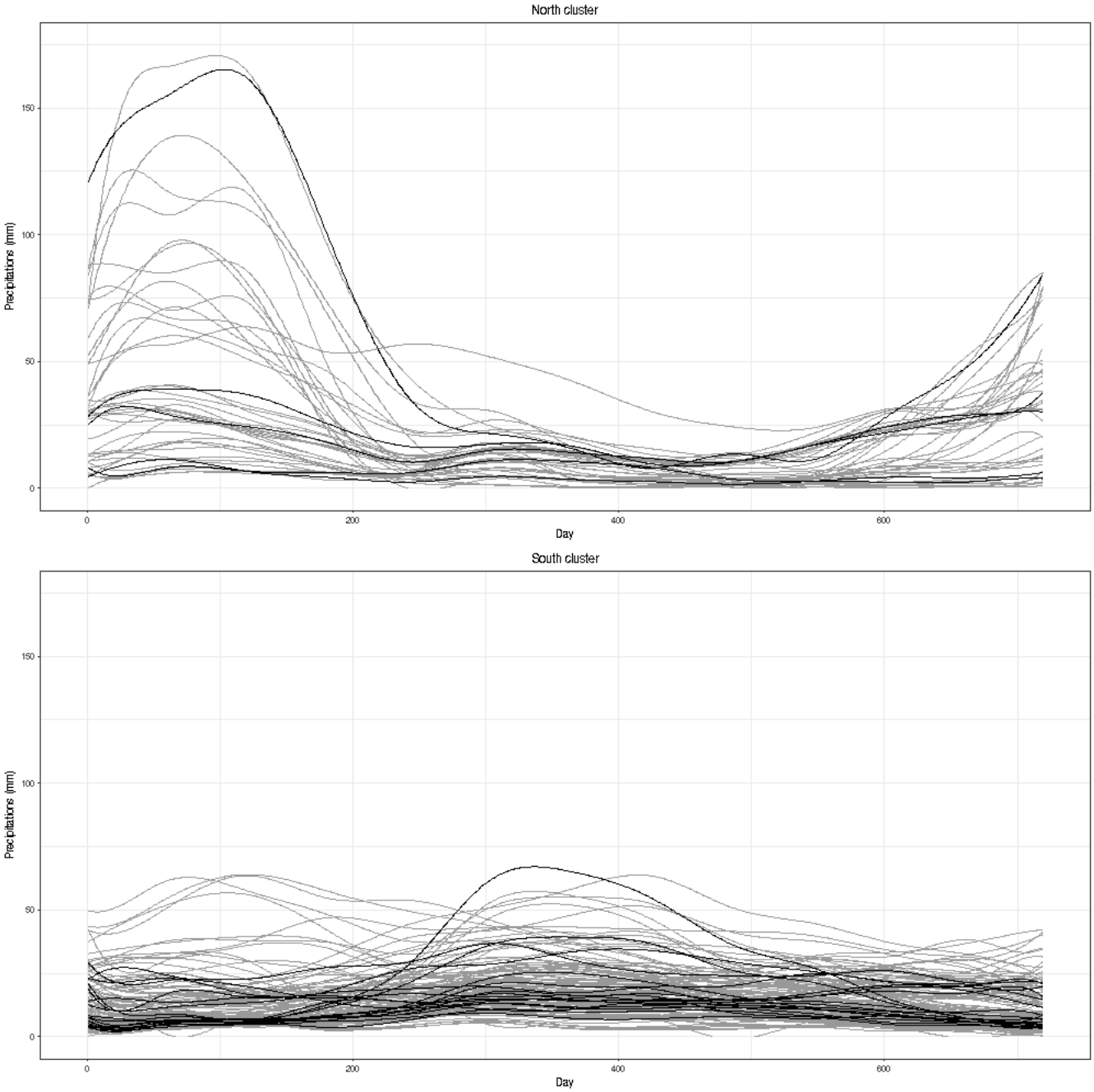

Figure 6: The North (top) to South (bottom) clustering of the Australian rainfall dataset with randomly chosen observations highlighted in black.

(black) clustering is highlighted on the left, where as the West (grey) to East (black) clustering is highlighted on the right.

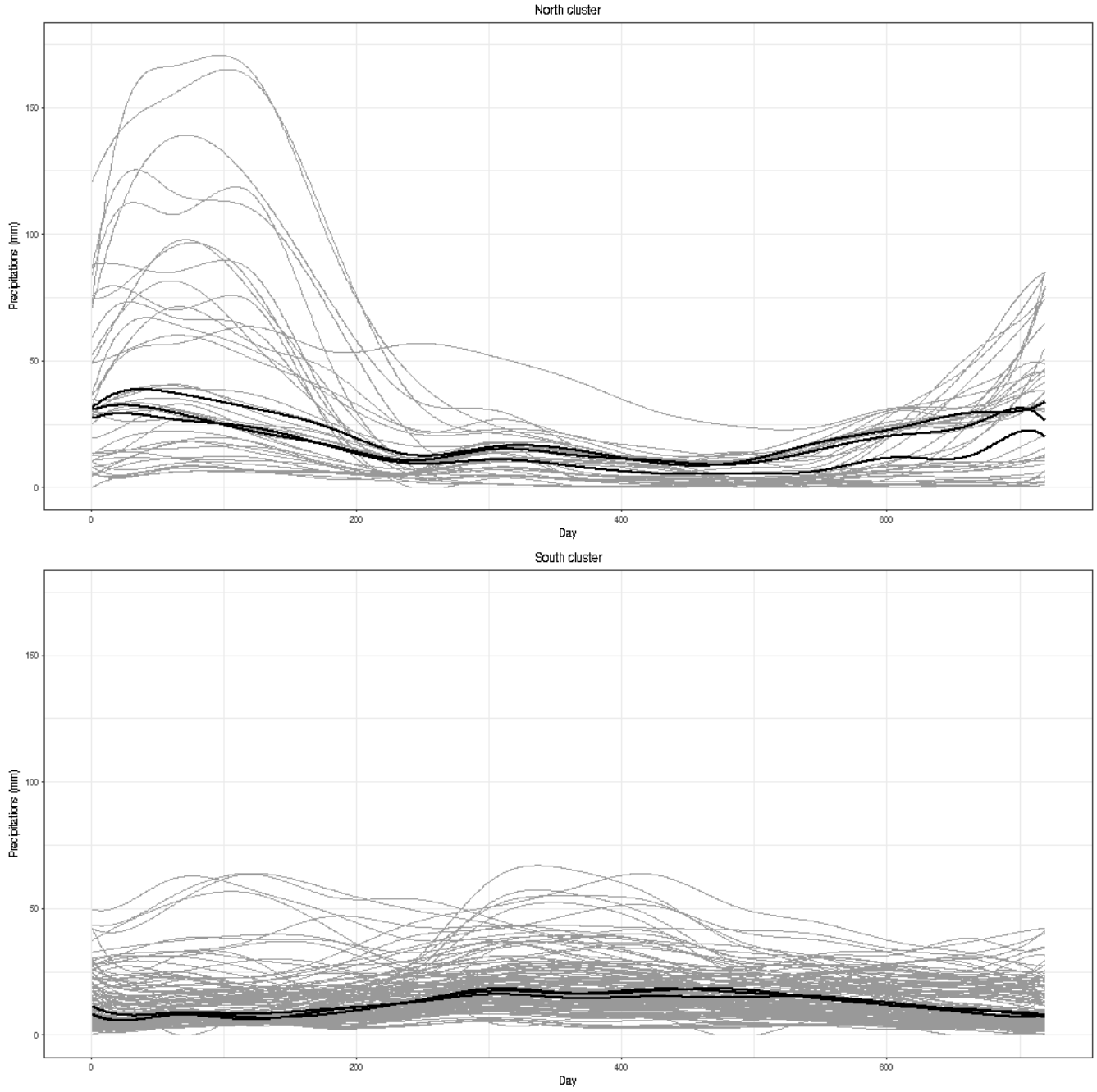

Figure 7: The North to South clustering with the three observations that obtained the highest depth values for AFD_1^2 highlighted in black.

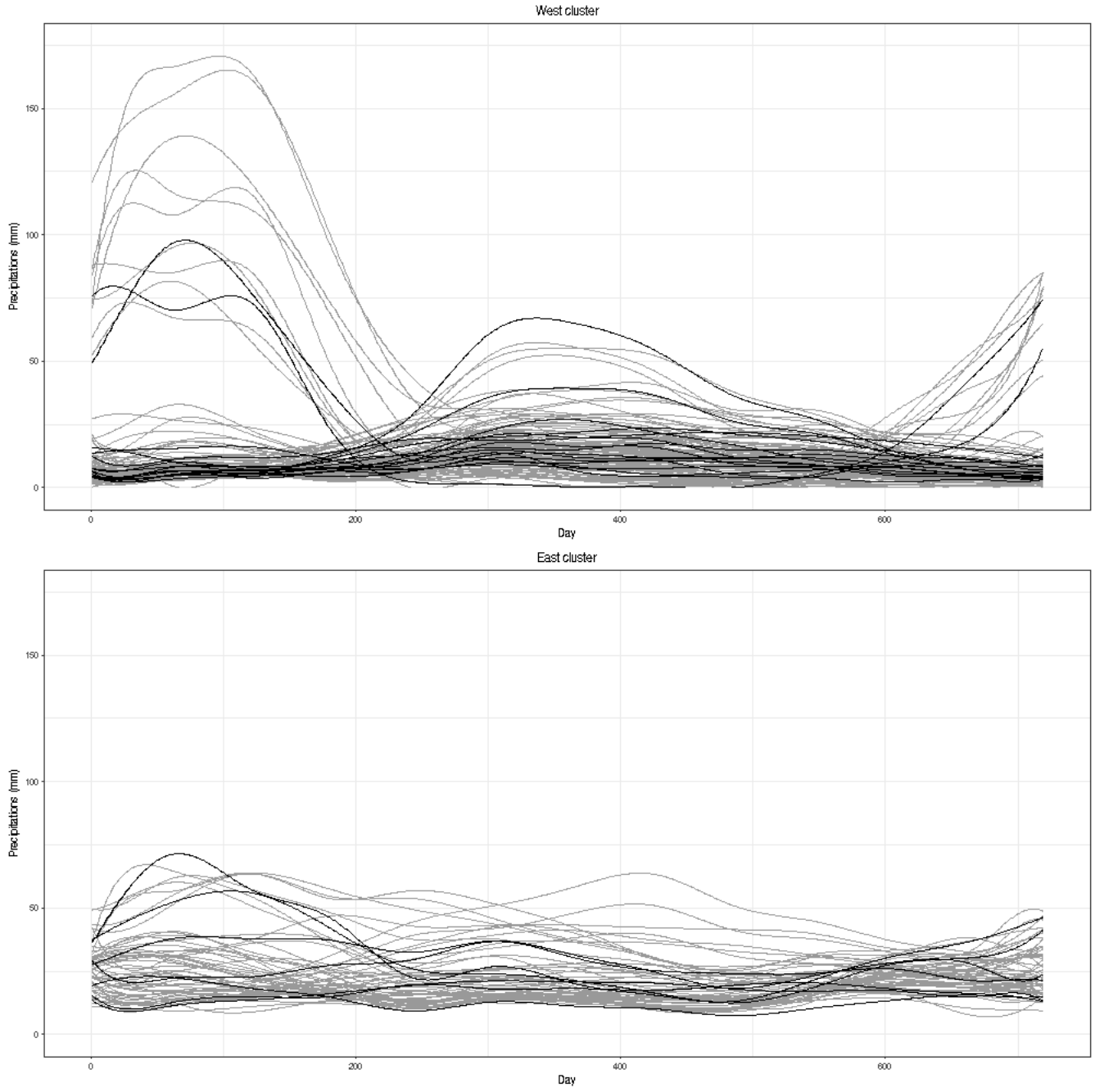

Figure 8: The West (top) to East (bottom) clustering of the Australian rainfall dataset with randomly chosen observations highlighted in black.

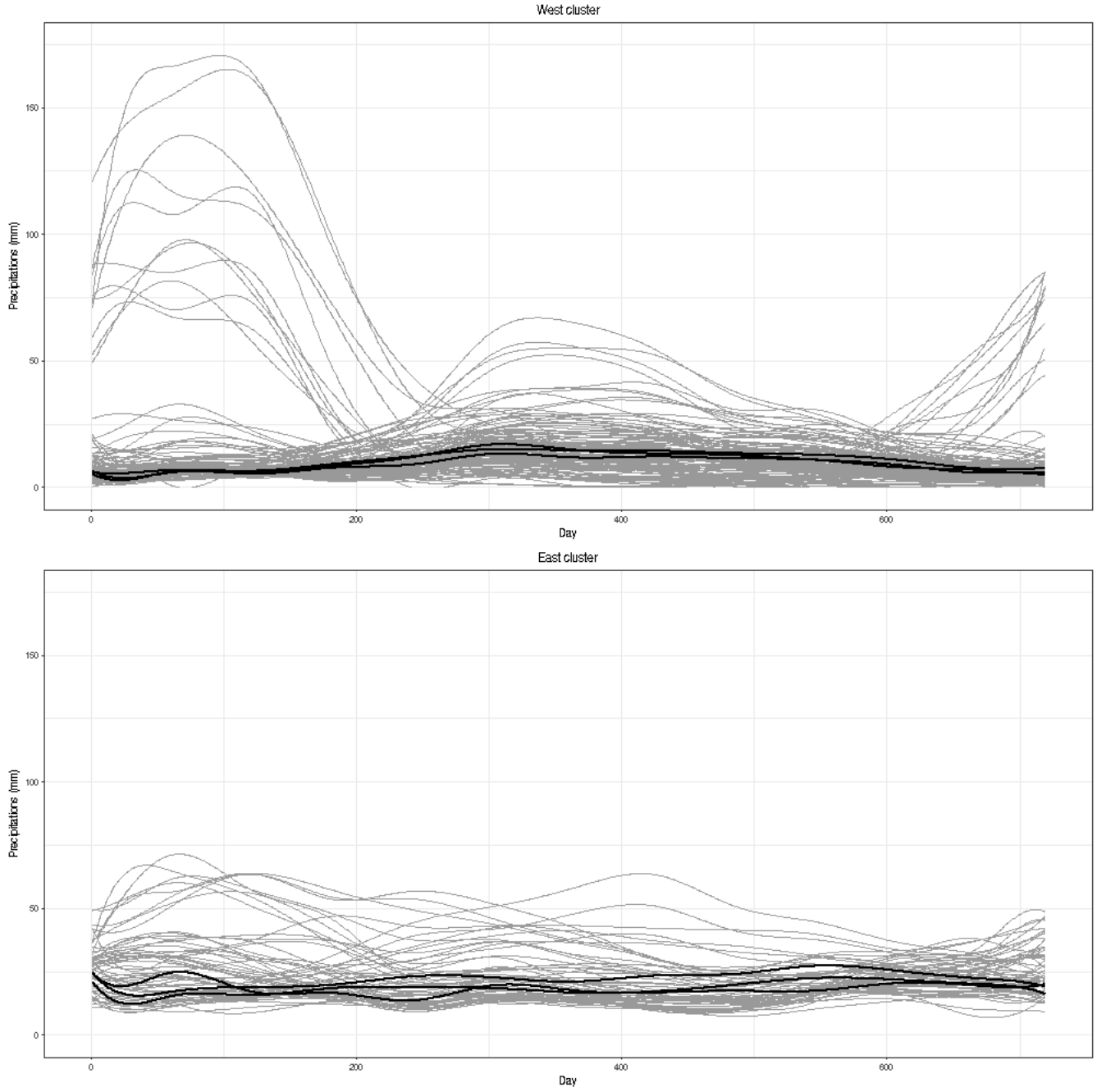

Figure 9: The West to East clustering with the three observations that obtained the highest depth values for AFD_1^2 highlighted in black.

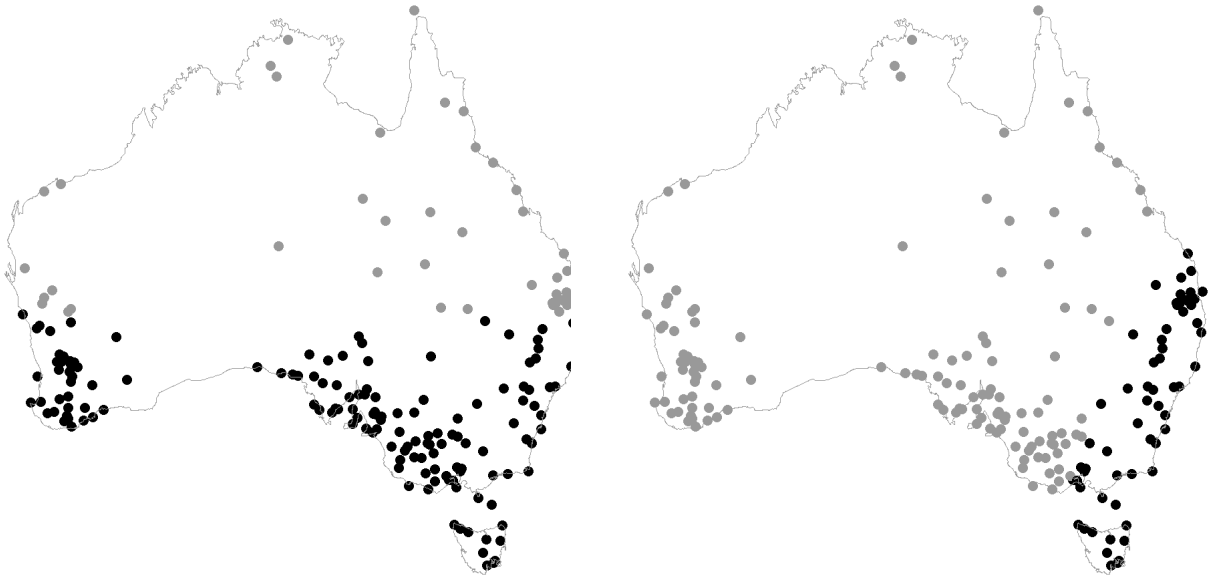

Figure 10: The geographical locations of the Australian weather stations with the NS (left) and WE (right) clusterings highlighted in grey and black.

The approximated Jth order kth moment depth AFD_k^J was computed using the orders $J = 1, 2$ and 3. For $J = 1$, the depths were computed across all possible time indices. For $J = 2$ and $J = 3$, the depths were computed over M_J uniformly sampled unique combinations of J time indices (with permuations removed), where $M_2 = 10\ 000$ and $M_3 = 5\ 000$. For each J, the moments $k = 1, 2$ and 3 were used.

The leave-one-out misclassification rates for AFD_k^J are presented in Table 1, and those for MFHD, KFSD and the PC and PLS classifiers, are presented in Table 2. For the WE cluster, both the competing depth methods as well as the PC and PLS classifiers perform similarly well, with the PLS having a slightly higher misclassification rate compared to the others. With any combination of the parameter J and k values, AFD_k^J clearly outperforms the other methods, aside from the PC classifier, compared to which it still performs slightly favourably. For the NS clustering, the performance of the depth methods, MFDH, KFSD and AFD_k^J for $J < 3$, decreases drastically, while the performance of the PC and PLS classifiers stays reliable. However, for $J = 3$, AFD_k^J is able to perform comparably, having only a slightly worse performance to the PC classifier.

Interestingly, the value of k did not seem to affect the performance of AFD_k^J for the WE clustering, but had a large impact in the case of the NS clustering. To confirm this, a leave-one-out crossvalidation scheme was conducted for the parameter k. For both classification problems, one at a time, each observation was left out of the pooled sample. Then, for the remainder of the sample, a leave-one-out classification was performed using a wide range of different values of k. Then, the value of k which had the lowest misclassification rate was used to classify the observation that was originally left out.

For the crossvalidation, the order J was chosen to be 1, and the values of k from between $-100\ 000$ and $100\ 000$ were tried. The value of k did not seem to have a meaningful effect on the misclassification rate for the WE clustering, and even with the crossvalidation, the misclassification rate was 6.3%. However, for the NS clustering, the crossvalidation had a significant impact as the misclassification rate fell down to 0.5% with only a single observation misclassified. Furthermore, the cluster from which the left out observation came from seemed to have a strong impact on the resulting crossvalidated value of k. This dependence on the left out observation is extremely rare for crossvalidation shcemes, inviting further exploration.

Table 1: Leave-one-out misclassification rates for WE and NS clusters for the Australian rainfall dataset based on within maximum depth classification with AFD_k^J.

	WE cluster	NS cluster
AFD_1^1	0.063	0.274
AFD_2^1	0.068	0.284
AFD_3^1	0.068	0.289
AFD_1^2	0.063	0.174
AFD_2^2	0.063	0.184
AFD_3^2	**0.058**	0.195
AFD_1^3	0.079	**0.132**
AFD_2^3	0.079	0.142
AFD_3^3	0.074	0.158

Table 2: Leave-one-out misclassification rates for WE and NS clusters for the Australian rainfall dataset based on within maximum depth classification with MFHD and KFSD, and based on the PC and PLS classifiers.

	MFHD	KFSD	PC	PLS
WE cluster	0.100	0.121	**0.084**	0.168
NS cluster	0.258	0.316	**0.084**	0.136

4.2 Kemijoki dataset

The Kemijoki dataset[2] (first analysed by Helander et al. (2018)) depicts the surface level of three different water reservoirs of mutually disconnected hydro power plants on the Kemijoki river. Each observation depicts the development of the reservoir surface level over a single day, and the data consists of observations from 484 different days drawn from multiple years. For each day, the measurements are given as a difference from the maximum reservoir level in meters, measured every 10 minutes resulting in observation sequences of 144 measurements for each day. For simplicity, these differences are henceforth referred to as "levels". The data is presented in Figure 11 with randomly chosen observations highlighted in solid black line to help perceive the overlapping observations.

Reservoirs A and C are very similar in both location and spread, but differ in shape. Their daily means (average level during a given day) are close to one another, with an average value of -0.23 and -0.24 respectively, and are very similarly spread aside from some clearly outlying observations. Aside from the slightly tighter grouping of the majority of the data of reservoir C, the notable differences between the two reservoirs are in the shape of the observations. Observations from reservoir A tend to have less surface level fluctuation and exhibit a slightly increasing trend until time 45-50, after which the amplitude and volatility of the fluctuations increase and the trend turns to stable or slighly decreasing. For reservoir C, the general trend is much more stable and the surface level fluctuations are much slower and less volatile, resulting in lower but more even roughness in the data compared to reservoir A. Reservoir B stands clearly separate from the other two in both location and shape. Its observations are located around a daily mean level of -0.15 and exhibit much less overlap than those of reservoirs A and C. The observations of reservoir B fluctuate much less over the course of a day, with the water level typically staying very stable for long periods of time compared to the rapidly fluctuating levels of the other two reservoirs. Furthermore, reservoir B is much less spread than the other reservoirs, having a standard deviation of its mean levels of only 0.035 , compared to that of 0.072 and 0.079 for reservoirs A and C respectively. The three observations that obtained the highest depth values with AFD_1^2 for each reservoir pair are highlighted in Figure 12.

Given the clear difference in both shape and location between reservoir B and reservoirs A and C, pairwise classification between B and either of the other two is relatively straightforward. However, due to the overlap of the reservoirs A and C and with the only notable differences between these reservoirs being in the general shape of the observations, classification between these two reservoirs is a difficult task and relies heavily on the shape receptivity of the depth function.

[2]The data is shared by Kemijoki Oy to the scientific community for academic research purposes by the original request of Department of Mathematics and Systems Analysis at Aalto University, Finland. Any other use of the data is not allowed. Due to possible competitive advantage reasons, any distinguishing information of the data, including the dates and specific reservoirs, have been removed. The data is not publicly available, but can be redistributed for research purposes on request.

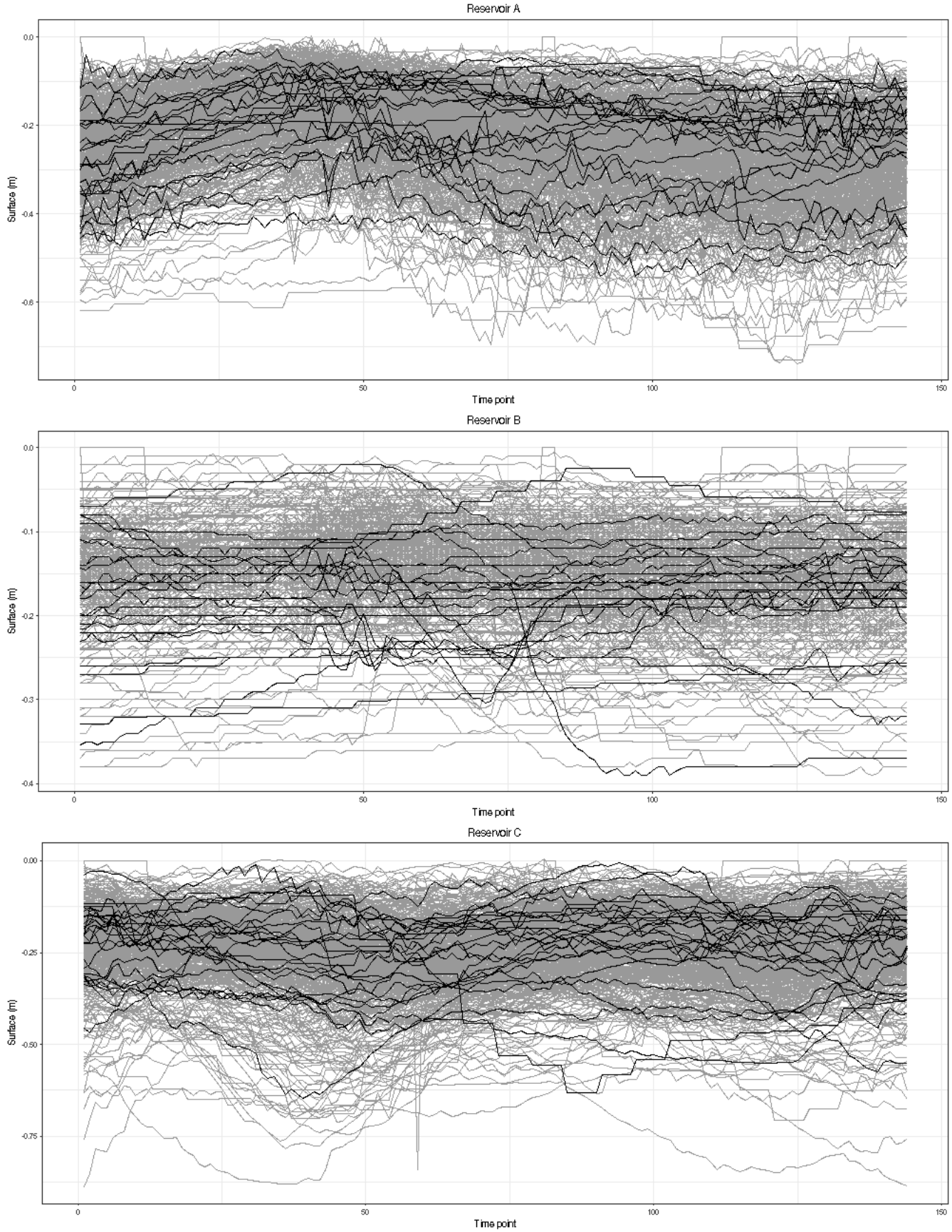

Figure 11: Kemijoki hydro power plant reservoir levels with randomly chosen observations highlighted in black.

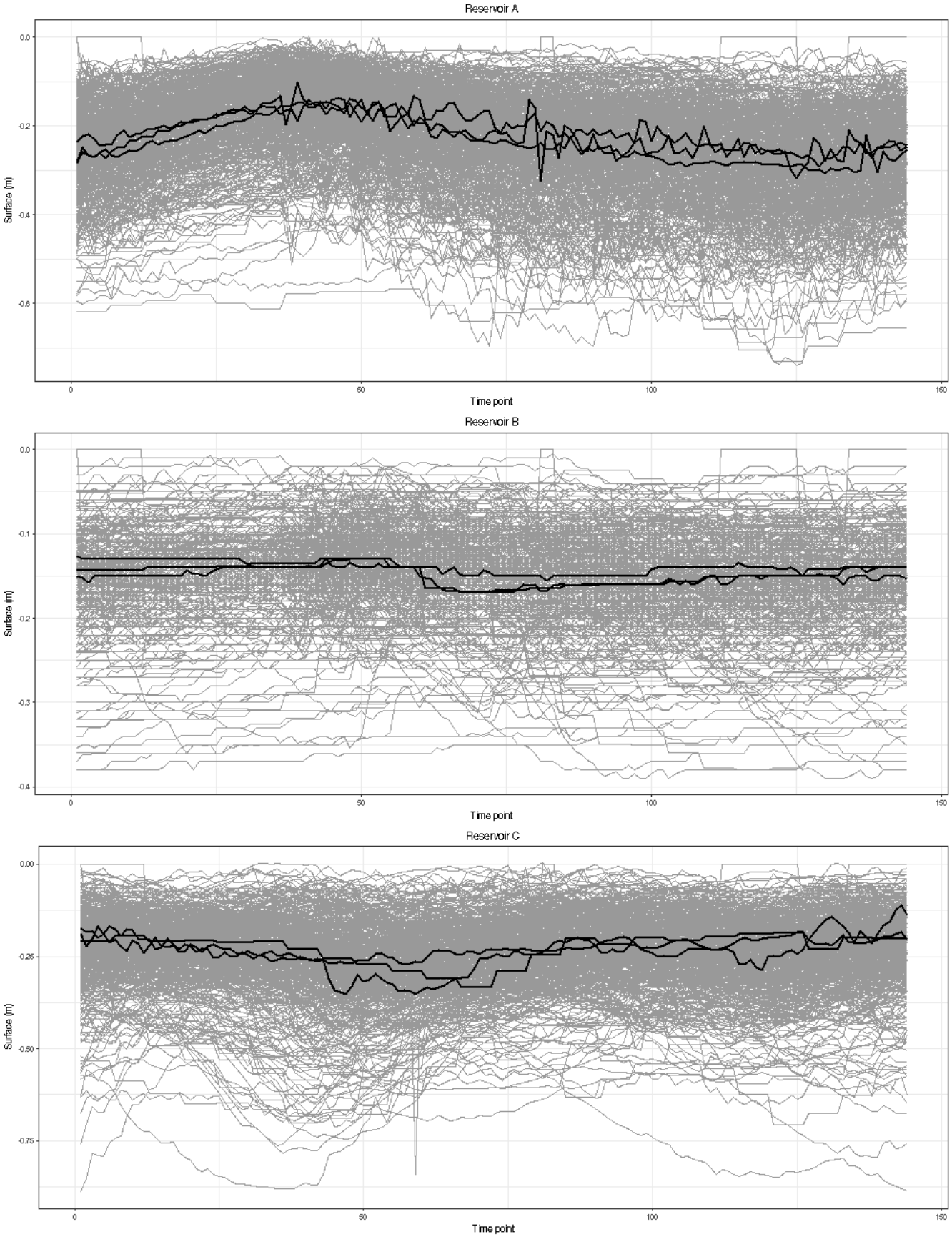

Figure 12: Kemijoki hydro powerplant reservoir levels with the three observations that obtained the highest depth values with AFD_1^2 for each reservoir highlighted in black.

The daily measurements were interpolated to form the functional observations using a B-spline basis of order 4 with knots placed at each measurement point. Thus, the resulting functional observations $x_i(t)$, $i = 1, \ldots, 484$ interpolate the data almost exactly, without removing the roughness.

For AFD_k^J, the moment $k = 1$ was held constant, and the orders of $J = 1, 2$ and 3 were used. With $J = 1$, the depth was computed across the entire index set, and for $J = 2$ all $\binom{144}{2}$ unique pairs of index points were used. With $J = 3$, the depths were computed over 5 000 uniformly sampled unique time index triplets, with no permutations.

The leave-one-out misclassification rates for all of the compared methods for the three reservoir pairings are presented in Table 3. The PC classifier clearly outperforms the other tried methods. The compared depth methods performed similarly well across each classification problem, with AFD_1^3 performing favourably. However, the weaker performance of the depth based methods, especially of AFD_k^J, compared to the Australian rainfall dataset may be explained by the roughness of the data. As the Kemijoki data was not smoothed, the present very volatile roughness in the data nearly drowns out the subtle differences in trend between the reservoirs interfering with the ability of the depth functions to capture the differences in shape.

Table 3: Leave-one-out misclassification rates for each reservoir pair based on within maximum depth classification with AFD_k^J, MFHD and KFSD and based on the PC and PLS classifiers.

	AFD_1^1	AFD_1^2	AFD_1^3	MFHD	KFSD	PC	PLS
AvB	0.236	0.217	0.198	0.229	0.249	**0.126**	0.234
BvC	0.204	0.201	0.192	0.214	0.207	**0.157**	0.200
CvA	0.325	0.256	0.240	0.262	0.243	**0.182**	0.266

5 Summary

In this thesis work we considered depth based classification of functionl observations. Most depth based methods presented in the literature fail to recognize shape features important to functional data. Our goal was to address this problem. That is why we concentrated on shape receptive depth functionals. In particular, we focused on the Jth order kth moment integrated depth and its application in supervised functional classificaiton using within maximum depth procedure. We presented the method, considered its properties, and illustrated its excellent performance in two different real data examples.

The Jth order kth moment integrated depth is based on integrating over the kth moments of J-variate crossectional depths. This approach enables to measure both shape and location even in the case when the derivatives of the functions do not exist. In the sample version, the integration is replaced by sums.

The method is computationally intensive, especially if J is chosen to be large. Thus, in the future, our goal is to provide a computationally more efficient algorithm (based on sampling randomly but wisely). Also, we plan to present a modified version of the Jth order kth moment integrated depth, where we consider not only J-variate cross sectional depths, but also simultaneously all j-variate cross sectional depths, with $j = 1, \ldots, J$. Moreover, we plan to apply Jth order kth moment integrated depths in clustering.

References

[1] Bartoszyński, R., Pearl, D. K. and Lawrence, J. (1997) A multidimensional goodness-of-fit test based on interpoint distances. *Journal of American Statistical Association*, **92**, 577-586.

[2] Biau, G., Bunea, F., and Wegkamp, M. (2005) Functional classification in Hilbert spaces. *IEEE Trensactions on Information Theory*, **51**, 2163-2172.

[3] Cérou, F. and Guyader, A. (2006) Nearest neighbor classification in infinite dimension. *ESAIM. Probability and Statistics*, **10**, 340-355.

[4] Chakraborty, A. and Chaudhuri, P. (2014a) On data depth in infinite dimansional spaces. *Annals of the Institute of Statistical Mathematics*, **66**, 303-324.

[5] Chakraborty, A. and Chaudhuri, P. (2014b) The spatial distribution in infinite dimensional spaces and related quantiles and depths. *Annals of Statistics*, **42**, 1203-1231.

[6] Chaudhuri, P. (1996) On a gemoetric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, **91**, 862-872.

[7] Claeskens, G., Hubert, M., Slaets, L. and Vakili, K. (2014) Multivariate functional halfspace depth. *Journal of American Statistical Association*, **109**, 411-423.

[8] Cuevas, A., Feberero, M. and Fraiman, R. (2007) Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, **22**, 481-496.

[9] Cuevas, A. and Fraiman, R. (2009) On dpeth measures and dual statistics. A methodology for dealing with general data. *Journal of Multivariate Analysis*, **100**, 753-766.

[10] Delaigle, A. and Hall, P. (2012) Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society, Series B*, **74**, 267-286.

[11] Durrett, R. (2010) Probability: theory and examples. *Cambridge university press.*

[12] Dutta, S., Ghosh, A. K. and Chaudhuri, P. (2011) Some intriguing properties of Tukey's half-space depth. *Bernoulli*, **17**, 1420-1434.

[13] Epifano, I. (2008) Shape descriptors for classification of functional data. *Technometrics*, **50**, 284-294.

[14] Ferraty, F. and Vieu, P. (2003) Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis*, **44**, 161-173.

[15] Fraiman, R. and Muniz, G. (2001) Trimmed means for functional data. *Test*, **10**, 419-440.

[16] Ghosh, A. K. and Chaudhuri, P. (2005) On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, **32**, 327-350.

[17] Gijbels, I., and Nagy, S. (2017) On a General Definition of Depth for Functional Data. *Statistical Science*, **32(4)**, 630-639.

[18] Hall, P. Poskitt, D. and Presnell, B. (2001) A Functional data-analytic approach to signal discrimination. *Technometrics*, **43**, 1-9.

[19] Helander, S., Van Bever, G., Rantala, S., and Ilmonen, P. (2018) Pareto Depth for Functional Data. *Submitted*

[20] Horváth, L., and Kokoszka, P. (2012) Inference for functional data with applications (Vol. 200). *Springer Science & Business Media.*

[21] James, G. and Hastie, T. (2001) Functional Linear Discriminant Analysis for Irregularly Sampled Curves. *Journal of the Royal Statistical Society, Series B*, **63**, 533-550.

[22] Lavery, B., Kariko, A. and Nicholls, N. (1992) A historical rainfall data set for Australia. *Australian Meteorological Magazine*, **40(1992)**, 33-39.

[23] Liu, R. Y. (1990) On a notion of data depth based on random simplices. *Annals of Statistics*, **18**, 405-414.

[24] Liu, R. Y. and Singh, K. (1993) A quality index based on data depth and multivariate rank tests. *Journal of American Statistical Association*, **88**, 252-260.

[25] Liu, R., Parelius, J. M. and Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *Annals of Statistics*, **27**, 783-858.

[26] López-Pintado, S. and Romo, J. (2006) Depth based classification for functional data. in *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, eds. Liu. R, Serfling, R., and Souvaine, D. L., Provedence: American Mathematical Society, DIMACS Series, pp. 103-120.

[27] López-Pintado, S. and Romo, J. (2009) On the concept of depth for funcional data. *Journal of the American Statistical Association*, **104**, 718-734.

[28] López-Pintado, S. and Romo, J. (2011) A half-region depth for functional data. *Computational Statistics and Data Analysis*, **55**, 1679-1695.

[29] Mahalanobis, P. C. (1936) On the generalized distance in statistics. *National Institute of Science of India*, **12**, 49-55.

[30] Marx, B. and Eilers, P. (1999) Generalized Linear Regression on Sampled Signals and Curves: a P-Spline Approach. *Tehcnometrics*, **41**, 1-13.

[31] Mosler, K. (2013) Depth statistics. In *Robustness and complex data structures*, pages 17-34, Springer, Heidelberg.

[32] Nagy, S., Gijbels, I., and Hlubinka, D. (2017) Depth-Based Recognition of Shape Outlying Functions. *Journal of Computational and Graphical Statistics*, **4**, 883-893.

[33] Nagy, S., Helander, S., Van Bever, G., Viitasaari, L., and Ilmonen, P. (2018) Depth-moments in nonparametric classification of functional data. *Manuscript*.

[34] Nieto-Reyes, A. (2011) On the properties of functional depth. In *Recent advances in Functional Data Analysis and Related Topics: Selected papers from the 2nd International Workshop on Functional and Operational Statistics (IWFOS'2011)* (F. Ferraty, ed.) Physica-Verlag / Springer, Heidelberg, 239-244

[35] Nieto-Reyes, A. and Battey, H. (2016) A topologically valid definition of depth for functional data. *Statistical Science*, **31**, 61-79.

[36] Paindaveine, D. and Van Bever, G. (2012) Nonparametrically consistent depth-based classifiers. *Bernoulli*, **21(1)**, 62-82.

[37] Paindaveine, D. and Van Bever, G. (2013) From Depth to Local Depth: A Focus on Centrality. *Journal of the American Statistical Association*, **108(503)**, 1105-1119.

[38] Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis, 2nd Ed.* Springer-Verlag.

[39] Rousseeuw, P. J. and Ruts, I. (1996) Bivariate location depth. *Royal Statistical Society Series C*, **45**, 516-526.

[40] Rousseeuw, P. J. and Struyf ,A. (1998) Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, **8**, 193-203.

[41] Rousseeuw, P. J. and Hubert, M. (1999) Regression depth (with discussion). *Journal of American Statistical Association*, **94**, 388-433.

[42] Ruts, I. and Rousseeuw, P. J. (1996) Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis*, **23**, 153-168.

[43] Serfling, R. (2002) A Depth function and a scale curve based on spatial quantiles. *Statistical Data Analysis Based on the L1-Norm and Related Methods*, ed. Didge, Y., Basel: Birkhauster, pp. 25-38.

[44] Serfling, R. (2010) Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *Journal of Nonparametric Statistics*, **22:7**, 915-936.

[45] Sguera, C., Galeano, P. and Lillo, R. (2016) Functional outlier detection by a localdepth with application to NOx levels. *Stochastic Environmental Research and Risk Assessment*,**30**, 1115–1130.

[46] Tukey, J. W. (1975) Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. V., 1974), Vol. 2*, 523-531 . Canadian Mathematical Congress, Montreal, Que.

[47] Vardi, Y. and Zhang, C.-H. (1999) The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, **97(4)**, 1423-1426.

[48] Koshevoy, G. and Mosler, K. (1997) Zonoid trimming for multivariate distributions. *Annals of Statistics*, **25**, 1998-2017.

[49] Zuo, Y. and Serfling, R. (2000a) General notions of statistical depth function. *Annals of Statistics*, **28**, 461-482.

[50] Zuo, Y. and Serfling, R. (2000b) Nonparametric notions of multivariate "scatter measure" and "more scattered" based on statistical depth functions. *Journal of Multivariate Analysis*, **75**, 62-78.